THE UNIVERSITY
of EDINBURGH

Evaluating the utility of gene expression data from patient-matched samples for studying breast cancer.

Richard Bownes

Doctor of Philosophy
The University of Edinburgh
2019

Declaration

I, Richard Joshua Bownes, confirm that the work presented in this thesis is my own and has not been submitted elsewhere for any other degree or professional qualification. Where information has been derived from other sources or in collaboration with colleagues, I confirm that this has been indicated.

Abstract

Breast cancer is a heterogeneous disease with distinct subtypes and many different clinical presentations. Neoadjuvant therapy of breast cancer offers a window of opportunity to study translational changes in tumours as a result of treatment alone and may help to identify tumour response status. Pairs of samples collected from different sites or sequentially from the same individual can potentially provide additional prognostic information for the risk stratification of breast cancer. Here, we seek to aggregate multiple studies of valuable, multi-sampled, patient-matched cohorts for meta-analysis to check for an enhanced ability to make new and significant findings about the underlying mechanisms of tumour treatment response.

Multiple sequentially-matched datasets of pre- and on-treatment matched primary tumour and lymph node samples were collected and examined for differentially expressed genes and pathways indicative of pathological response. Machine learning methods were applied to identify biomarkers of response from the on-treatment samples, and profiling comparisons were made to assess the additional value of matched patient samples to accurately predict risk. Lastly, five sequentially sampled datasets were aggregated for meta-analysis by combining the normalised pre- to on-treatment expression level differences to identify commonalities in the response to therapy across both endocrine and chemotherapy treatment strategies.

The gene, AAGAB, was identified through iterative differential analysis, and was found to be 78% accurate in validation for the prediction of pathological complete response in neoadjuvant chemotherapy treated breast cancer. AAGAB demonstrated significant separation of patient survival curves (log rank $p = 0.0036$), and the on-treatment samples more accurately reflected the patient risk than the pretreatment samples. Matched lymph node tissue of primary breast cancer was more successful at capturing the patient's risk of recurrence than the primary biopsy, correctly identifying 83% (10/12) of the recurring patients compared to 25% (3/12) in the primary. Underlying differential expression analysis also showed a considerable number of high profile breast cancer genes over-represented in the lymph node. Aggregation of multiple sequential studies resulted in low post integration concordance values with the reference patient data (<30% profiling agreement), and is not recommended for this type of analysis. However, combining the pairwise change values for gene expression level data was successful, and resulted in the creation of highly accurate models for predicting patient response (F1 accuracy score, 0.92) as well as the identification of

potential common escape pathways to breast cancer therapies.

Analysis of the matched pre- and on-treatment samples revealed the intrinsic value of multiple on-treatment biopsies. These samples offer valuable new targets for biomarker identification that show significant increases in accuracy for the prediction of response and long term outcome in neoadjuvant chemotherapy. Additional sampling of involved metastatic lymph node also improves the prognostic capabilities for clinicians by providing a potentially more accurate view of the per-patient risk profile. Lastly, the pairwise expression change values show the direction of tumour change, which can be used to create new models for the prediction and classification of patient risk and for furthering our understanding of the mechanisms behind patient non-response.

Lay-Abstract

Breast cancer is not one disease but many, and there is evidence that each subtype has distinctly different characteristics. For many years, the standard of care for breast cancer was surgery followed by radiation therapy or more targeted treatments, if deemed appropriate. A possible alternative approach is for patients to receive treatment prior to surgery. This method shows promise for distinguishing patients with good outcomes from patients with poor outcomes. Pre-surgery treatment was recently shown to be as effective as the standard care treatments, but may improve our ability to identify new ways to detect response to treatment in patients and help future patients.

The most common site for a tumour to move to, and the first place breast cancer usually spreads, is to the nearby lymph nodes of the armpit. When tumour cells are detected in the lymph node, patients are known to have worse outcomes and are more likely to have a future reappearance of cancer. At present, when the tumour is present in lymph nodes, this information is not used beyond a categorical risk factor for the patient. However, the progression to the lymph nodes and the profile of that tumour may hold important information for new treatment parameters for each individual. It is possible that the characteristics inherent to the lymph node tumours that facilitate the progression from the breast allow these samples to better represent the risk of the patient as a whole.

As pre- and on-treatment matched samples and patient matched tissue samples are both rare, analysis of these samples is often of limited value. Combining available data for analysis would boost the significance of all findings derived from this data. However, integration of this data must be undertaken carefully and with enough subtlety to not erase the underlying biological information. This proves a non-trivial task but has a large potential payout for future analysis and biomarker testing.

These informative matched paired samples have helped identify new methods for patient risk prediction and combining many smaller studies has improved our understanding of the differences between different types of patients and treatment strategies.

Acknowledgements

This thesis might contain all my own words, but it isn't the product of only one person. I need to thank my supervisor, Andy Sims, for his constant supervision, patient and helpful oversight, guidance and experience. I would not have been able to succeed these last few years with a less capable or less kind mentor. Dr. Olga Oikonomidou for co-supervision and for granting access to the NEO trial, which comprised a major portion of my thesis work and all of the patients whose samples made all of my work possible. I would also like to acknowledge my funding body, CRUK, for their generosity.

I would also like to thank my family for their invaluable support. My parents are my inspiration in life and have provided help, support and advice beyond what I can write. During my PhD, I adopted two dogs, who were the missing limbs I didn't know I didn't have. Bean and Isla have taught me responsibility in a way I never knew before. They also helped to get me outside every once in a while. Lastly, my best friend and the most amazing woman I know, thank you Max, this would not have happened without you. You have helped my growth as a student and an academic, you are responsible for my growth as a person, and you are my reason to keep to moving. Check it out, we did it!

In light of recent events and with the global pandemic family and community have never been more important. It is with a heavy heart that this thesis is now being resubmitted under the current circumstances. Like it takes a village to raise a child it takes a tribe to produce a body of work like this.

We who cut mere stones
must always be envisioning cathedrals.
    - Stone cutters's creed

Contents

Abbreviations

AAGAB	**A**lpha **A**nd **G**amma **A**daptin **B**inding-Protein
AI	**A**romatase **I**nhibitor
API	**A**pplication **P**rogramming **I**nterface
AURKA	**AUR**ora **K**inase **A**
BC	**B**reast **C**ancer
BRCA	**BR**east **CA**ncer
CB	**C**ore-needle **B**iopsy
CDK	**C D K**
CMF	**C M F**
cfDNA	**c**ell-**f**ree **DNA**
ctDNA	**c**irculating **t**umour **DNA**
DGEA	**D**ifferential **G**ene **E**xpression **A**nalysis
DNA	**D**eoxyribo **N**ucleic **Acid**
EB	**E**xcision **B**iopsy
EGFR	**E**pidermal **G**rowth **F**actor **R**eceptor
ER	**E**strogen **R**eceptor
FF	**F**resh **F**rozen
FFPE	**F**ormalin **F**ixed **P**arafin **E**mbedded
GSEA	**G**eneset **E**nrichment **A**nalysis
HER2	**H**uman **E**pidermal-Growth-Factor **R**eceptor
HR	**H**ormone **R**eceptor
IGF	**I**nsulin-like **G**rowth **F**actor
IHC	**I**mmunohistochemistry **C**hemical
JSON	**J**ava**S**cript **O**bject **N**otation
LumA	Luminal A
LumB	Luminal B
LFDA	**L**ocal **F**isher **D**iscriminant **A**nalysis
MAPK	**M**itogen **A**ctivated **P**rotein **K**inase
METABRIC	**M**olecular **T**axonomy **O**f **B**reast **C**ancer **I**nternational **C**onsortium
MDS	**M**ulti **D**imensional **S**caling
NAC	**N**eo**a**djuvant **C**hemotherapy

NCBIGEO	**N**ational **C**entre for **B**iotechnology **I**nformation **G**ene **E**xpression **O**mnibus
NIT	**N**o **I**ntervening **T**reatment
NKI	**N**ational **K**aker **I**nstituut
NPI	**N**ottingham **P**rognostic **I**ndex
OR	**O**dds **R**atio
PCA	**P**rincipal **C**omponent **A**nalysis
PCNA	**P**roliferating **C**hain **N**uclear **A**ntigen
PCR	**P**athological **C**omplete **R**esponse
POETIC	**P**eri **O**perative **E**ndocrine **T**herapy for **I**ndividualising **C**are
PR	**P**rogesterone **R**eceptor
RNA	**R**ibonucleic **A**cid
ROR	**R**isk **O**f **R**ecurrence
SAM	**S**ignificance **A**nalysis of **M**icroarrays
T2	**T**ime **2**nd
TM	**T**ime **M**id-chemo
TNBC	**T**riple **N**egative **B**reast **C**ancer
TP	**T**ime **P**re-treatment
TS	**T**ime **S**urgical
tSNE	**t** distributed **S**tochastic **N**eighbour **E**mbedding
USS	**U**ltra **S**ound **S**onography

1 | Introduction

1.1 Cancer and Therapy

Cancer represents a state of fundamental dysregulation of the biological and cellular processes that underpin survival and fitness in healthy cells. Common hallmarks of cancer include growth signalling pathways, apoptosis, and replication;[1] all of which represent control over the cell cycle. In this way, cancer is a corruption of otherwise normal, healthy cells generating malignant cells which form clusters called tumours. Therefore, cancer therapies treat the tumour to the detriment of the body, and must seek to target characteristics inherent to the process of tumourigenesis, such as the significantly higher rates of proliferation and damaged DNA repair mechanisms.[2] Anti-mitotic agents that target DNA replication mechanisms are effective at targeting tumour cells, but are systemic and non-specific, placing a significant burden on the patient.[3] Thus, reducing over-treatment and identifying patients who will have good responses to therapy is an important area of research with real clinical impact[i] and the potential to improve patient care.

1.2 Breast Cancer - Incidence and Outcomes

Breast cancer (BC) is the most prevalent new cancer[ii] among women and the second most deadly.[4–6] Globally, it is the most common cancer in women, with 1.7 million new cases reported each year and over 500,000 deaths in the year 2012.[7] Alone, it accounts for a quarter of all cancer cases and 15% of cancer deaths in women.[7] Breast cancer represents a significant number of the total new cases of cancer every year in the UK, with 54,724 new cases in 2017 (or 15% of all new cancer cases) and a total mortality of 11,433.[8] Thanks to advances in the treatment of BC in the last several decades, the total mortality rate has dropped 32%

[i] Due to my unique position between clinical oncology and dry lab research, I only mean to emphasise the clear target of producing said results.

[ii] Specifically malignant cancer, ductal carcinoma *in situ* is present in an overwhelming number of women according to necropsy results, but is largely asymptomatic.

since the 1970's and 17% in the last decade.[8] Early detection of BC can help to increase the survival rates in patients by significant margins, as much as 20% after 5 years.[9] Currently, breast cancer has an 89.7% 5 year-survival rate[10] and a 78% 10-year survival rate in the UK (female cancers only).[8] Improvements in screening, treatment and diagnosis have produced a positive trend in the survival rate of BC and are continued fields of research.[11]

In 2012, there were 1.7 million newly diagnosed cases and 500,000 deaths from BC globally, representing almost a quarter of all newly diagnosed cancers and 15% of cancer related deaths in women.[12] These numbers represent a heterogeneous disease of different clinical, histological and intrinsic factors.[13] In supervised analysis of patients with known clinical and prognostic outcomes, and despite tumour cellular heterogeneity, inter-patient variation often exceeds the variance between groups of patients with regard to biological or clinical status.[14] This means identifying trends that significantly differentiate translational changes between responsive and non-responsive BC can be very difficult.

1.3 Detection, Diagnosis and Subtypes

In the UK, all women between the ages of 50 and 70 are invited for a breast screen every 3 years. This results in approximately 2 million women a year undergoing breas screening in the UK.[15] These screenings are responsible for reducing mortality in the UK by as many as 1,300 women per year. Additionally, breast masses can be detected by a physician, usually following a clinical breast examination as a result of patient reported symptoms.[15,16] A lump can be ratified by mammogram, MRI and ultrasound, but a definitive diagnosis can only be given through biopsy. Biopsies are performed as either a fine needle aspiration biopsy, a core needle biopsy or as a surgical biopsy, and the sample is examined by a pathologist to determine the malignant or benign nature of the mass. The pathology report contains key information that along with clinical information will facilitate treatment decision making and will also offer important information about the patient's prognosis.[17,18]

Breast cancer is broadly defined as a malignancy originating in breast tissue of men or women,[19] and is frequently described by the structure where it originates, either ductal or lobular or as one of a few rarer sub-categories of BC, including Paget's, Phyllodes or inflammatory BC.[20] BC is then more specifically defined as either *in situ* (in its original place) or invasive, and both are fur-

ther dichotomised into more refined subtypes, usually ductal or lobular,[18]see Figure 1.3.1 on Page 3. Ductal carcinoma *in situ* (DCIS) is the most common *in situ* subtype of BC, and is further described as either comedo or non-comedo, where the latter are further classified as cribiform, micropapillary, papillary, and solid.[18,21]

Invasive carcinomas are those that have infiltrated past their point of origin and are subdivided into different groups based on their histological features, the most common being Invasive Ductal Carcinoma (IDC) of no special type, which accounts for between 70 and 80% of all invasive BC tumours.[22]

Breast cancer tumours are typically divided into three grades as a function of its differentiation; grade 1 (well differentiated), grade 2 (moderately), and grade 3 (poorly).[22] Other common categories of invasive carcinoma include lobular, ductal/lobular, mucinous, medullary, papillary and tubular.[18] See Figure 1.3.1 on Page 3 for an illustration of the clinical subtypes of BC.

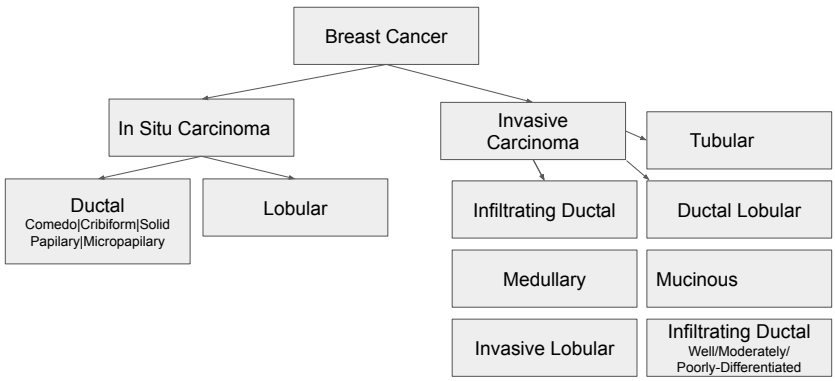


Figure 1.3.1: Histological Subtypes of Breast Cancer. An illustration of the broad categories of histological BC subtypes.

HER2-enriched tumours are those with amplification/over-expression of HER2 (Human Epithelial growth factor Receptor 2) and, which under IHC (Immunohistochemical) classification, appear as ER-(Oestrogen Receptor)/PR-(Progesterone Receptor)/HER2+.[23] HER2-enriched tumours frequently overexpress related genes to HER2, namely GRB7 (Growth Factor Receptor Bound Protein 7) and PGAP3 (Post-GPI Attachment To Proteins Phospholipase 3) and are most likely grade 3.[24] Additionally, these tumours often have poor prognoses,[25]

but have good rates of pathological clinical response due to the availability of targeted agents and susceptibility to standard chemotherapies.[26]

Basal tumours, or triple negative tumours, are characterised under IHC as ER-/PR-/HER2- and have expression profiles similar to the basal epithelial cells of breast tissue,[27] but are not necessarily identical. A summary of the characteristics of the molecular subtypes is included in Figure 1.3.2 on Page 5. Due to the lack of receptor targets, the only standard of care is chemotherapy, which, in conjunction with the generally aggressive nature of these tumours, leads to basal-like tumours having the worst clinical prognosis of the molecular subtypes.[24,28] The poor prognosis is also partly explained by the lack of treatment options available and recurrence rather than a lack of treatment efficacy.[26] On the other end of the response spectrum are ER+ tumours, the most common and of least severe general prognosis, are defined by their positive oestrogen receptor status. While oestrogen receptor negative tumours tend to be of higher grade and have a worse prognosis than ER+ counterparts, a small subset of triple negative BCs have better clinical outcomes and respond well to neoadjuvant chemotherapy, showing much better overall survival.[29,30] Further analysis has shown that TNBC (Triple Negative Breast Cancer) is a heterogeneous disease with between four and six potential subtypes, each characterised by distinct patterns of gene expression that have different treatment and survival characteristics.[29–31] This subsetting of TNBC may be important for future drug discovery and biomarker generation for these difficult to treat tumours.[31] BRCA (BReast CAncer gene) status is another important consideration for prognosis and treatment. BRCA gene mutations are highly associated with increased likelihood of developing specific epithelial cancers, most notably breast and ovarian.[32]

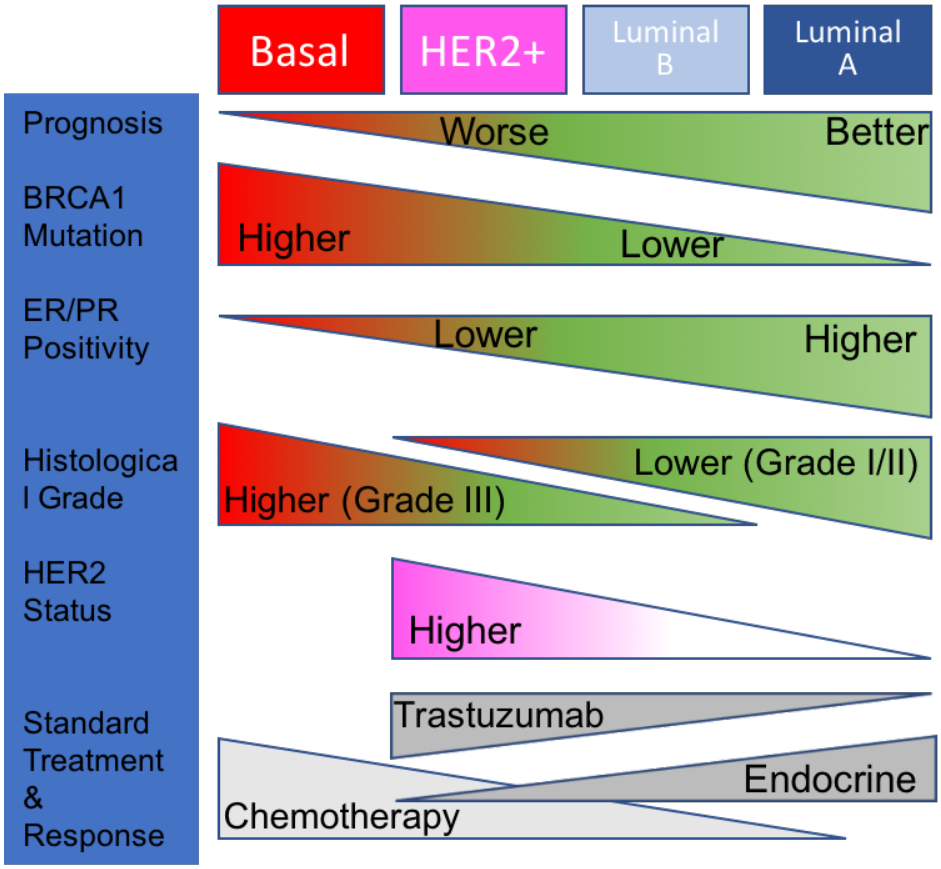


Figure 1.3.2: Breast Cancer Molecular Subtypes, Prognosis, Receptor Status and Treatment Options. This diagram shows the molecular subtypes of BC and how their receptor status relates to their relative prognosis.

1.4 Treatment Options and Decision-making

The most common and conventional means of treating BC are radiotherapy, chemotherapy, hormone therapy and targeted therapies.[33] Conservation of breast tissue in the treatment of localised disease, rather than mastectomy, is now the most common surgical approach to BC,[34,35] additionally it is the preferred surgical approach by both patients and clinicians. When surgery is preceded by treatment (neoadjuvant setting), it may shrink the tumor volume and reduce the amount of breast tissue resected.[36] Further treatment is usually given after surgery (adjuvant setting) to reduce the risk of recurrence or metastasis and improve patient outcomes.[33,36] Choosing a therapeutic pathway should be made with consideration of available information on the tumour, patient characteristics, pathological features, and histological information, such as the patients risk of recurrence and metastasis as well as the patient's wishes.[37] Additionally, the inclusion of intrinsic phenotypes such as ER/PR (Oestrogen/Progesterone Receptor), HER2 (Human Epithelial growth factor Receptor 2) and KI-67 (a gene marker of proliferation) should be considered in the treatment decision making process where possible at these factors have been shown to add prognostic value to clinicians.[38]

Non-specific treatments for BC are radiation therapy and chemotherapy. These treatment options do not discriminate between cell types but are designed to take advantage of cancer's proliferative nature. Radiation therapy is the direct exposure of the cancer cells to radiation, and is frequently given in combination with chemotherapy, as it can be used in early local cancer to reduce the chance of recurrence.[39] Chemotherapy is usually given to patients presenting as ER-, especially triple negative tumours (negative ER/PR/HER2 status), however it is also prescribed to HER2+ patients as well as Luminal tumours (ER+) at higher risk.[37] There are a few options in the treatment of cancer using chemotherapy. Historically CMF (cyclophosphamide, methotrexate, fluorouracil) was used in practice more than 15 years ago, but more recently taxanes and anthracyclines have been prescribed. Taxanes and anthracyclines may be given independently or in combination, and have been shown to have improved performance over treatment with CMF.[40] The combination therapy of taxanes and anthracyclines was shown to reduce mortality rates by one third, regardless of patient age or tumour characteristics.[41] However, these compounds are not generally well tolerated and can have adverse toxic effects, especially to those with preexisting co-morbidities and heart problems.[42]

Hormone, or endocrine, therapy (ET) is a targeted treatment option for

BC. The purpose of endocrine (hormone, ET) therapy is to block the activity of hormones and is recommended to all patients with significant endocrine receptor expression (defined as >1% of the cells).[33] ET can be used in addition to chemotherapy, radiation and targeted therapies, as long as the patients are ER+.[33] The choice of treatment, however, is dictated by the patient's menopausal status and individual risk.[33] For pre-menopausal patients, treatment with an ER antagonist is standard (e.g., tamoxifen, 5-10 years) and can be combined with ovarian ablation (surgically or chemically), but comes with associated risks of fertility and endometrial hyperplasia and cancer.[43] When combined with ovarian ablation, ET is at least as effective as treatment with CMF[44] and the combination of an AI (Aromatase Inhibitor) with surgical ovarian ablation or administration of a gonadotropin releasing hormone (GnRH) agonist is a tolerable substitute for tamoxifen.[45] In post-menopausal women, the ovaries are no longer producing oestrogen, therefore surgical ablation or chemical with GnRH-agonists are no longer necessary and the standard treatment for this cohort then becomes 5-10 years of adjuvant tamoxifen.[46]

Another prominent targeted therapy is trastuzumab which, in combination with chemotherapy, has been shown to reduce recurrence of HER2+ BC by half when compared to chemotherapy alone.[47] Further, improved treatment response was observed when trastuzumab was used in combination with chemotherapy alongside another HER2+ agent like Pertuzumab.[47,48] Trastuzumab has been approved for use in patients with positive nodal involvement or larger tumours (greater than 2 cm diameter) as well as any patient who will derive benefit from combined treatment and for concurrent use with taxanes, however, concurrent administration with anthracyclines is ill advised due to both therapies having notable cardiotoxicity.[49–51] Additional targets for targeted therapies include PI3K (Phosphoinositide 3-kinase)/mTOR (Mammalian Target of Rapamycin) and CDK (Cyclin-dependant kinase), as these are frequently dysregulated in BC.[52,53] Everolimus (an mTOR inhibitor) and CDK4/6 inhibitors (Palbociclib, Abemaciclib, Ribociclib) are FDA (Food and Drug Administration) approved drugs with improved treatment response in combination with existing chemotherapy agents in a variety of BC types. Common UK breast cancer treatment strategies are summarised in Figure 1.4.1 on Page 8, according to the European Society of Medical Oncology (ESMO).[54]

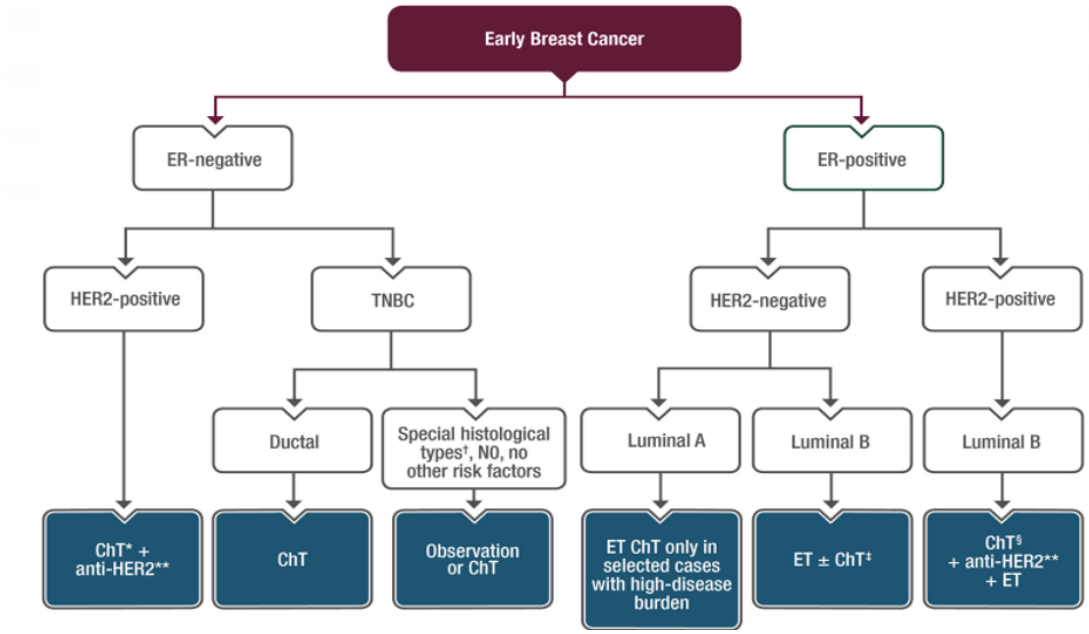


Figure 1.4.1: Generalised Treatment Strategies for Early BC. A basic overview of the treatment options for different clinically relevant subtypes of BC are shown in this diagram to illustrate the heterogeneous nature of BC treatment, sourced from ESMO.

1.5 Clinical Decision Making Tools

There are applied decision making tools currently in use in the UK that combine multiple clinical factors of the patient and tumour to help make informed treatment choices and prognosis. The Nottingham Prognostic Index (NPI) combines factors including the number of involved lymph nodes, tumour size, and tumour grade to calculate a score that is descriptive of prognosis in an attempt to identify high risk, poor prognosis patients.[55] The NPI has proven to be more informative than lymph node status alone as a means for predicting prognosis and has been validated in other British datasets and treatment centres.[55,56] Adjuvant! Online is another clinical tool that attempts to predict and quantify patient response to either endocrine or chemotherapy treatment in terms of predicted overall survival rates (patient survival from all causes of mortality) and recurrence free survival (survival based on no recurrence of the tumour).[57] These estimates are derived from patient age, menopausal status, tumour staging, number of involved nodes and ER status.[57] PREDICT NHS is another prognostic tool and the first that also encompasses mode of detection when calculating patient outcomes, and has been proven to be an effective and discriminatory means of prognostication.[56] In this regard, PREDICT NHS performs well across different prognostic groups and age ranges, and builds on work such as Adjuvant! Online

to create new and contemporary prognostic tools.[56]

1.6 Molecular Subtyping and Prognostic Signatures in Breast Cancer

In addition to histological subtypes, BC tumours can be described by receptor status and, more recently, by HER2 and TNBC (Triple Negative Breast Cancer) labels. Further subtyping, based on gene expression, has been popularised by Perou and Sorlie.[58] Though it should be noted that this molecular subtyping method is experimental. Molecular subtypes may be of use for predicting response to specific therapies[18] and, at least in the case of invasive carcinomas, it can be recommended to calculate the relevant sub-typing information for prognostic consideration.[59] These intrinsic subtypes of BC include Luminal A (High ER, low HER2) and Luminal B (Low ER, Low HER2 and High Proliferation), Her2 (HER2 positive, ER negative), Basal (ER negative, HER2 negative, PR negative) and normal-like.[58,60,61]

The generation of these subtypes may one day have widespread clinical applications for patient and clinician decision making because they have significantly different survival characteristics, for example, Basal like tumours having the poorest prognosis.[61] However, due to prohibitive costs, these methods have not seen broad clinical uptake, but the creation of targeted gene panels, like the FDA approved PAM50 (Prediction Analysis of Microarray 50), have made the generation of molecular subtypes possible on a small scale and have been shown to be comparable to full microarray analysis.[62] Methods like PAM50 have also been shown to out perform clinical information in predicting recurrence of BC. Importantly, utilising both targeted gene panels and clinical information such as T (Tumour size), N (Number of involved nodes), M (Metastatic status) and grade show improved predictive power, providing a potential future avenue for research.[62]

Seminal gene expression profiling studies by Sorlie and Perou described the five so-called “intrinsic” molecular subtypes of BC that correlate with distinct clinical applications in terms of treatment and prognosis.[25,27] Luminal tumours, A and B, have positive hormonal receptor status and have expression profiles similar to normal luminal epithelial cells of breast tissue, but differ more than “normal-type” tumours.[27] Immunohistochemically, they appear as ER+/PR+/HER2- (Luminal A) and ER+/PR+/HER2+/- (Luminal B),[23] but Luminal A have lower levels of expression for proliferation and are usually a lower grade

than Luminal B.[24,25] Luminal subtypes of breast cancer are the most common,[24] but have distinctly different prognoses; Luminal A usually exhibits a much better treatment outcome than Luminal B.[63] ER- tumours, of which there are two, make up half of the four distinct intrinsic molecular subtypes: Luminal A, Luminal B, HER2-enriched and basal-like.[27] The fifth subtype, normal-like, is now believed to be caused by contamination of normal mammary cells in the tumour biopsy.[64] Separating patients into these subtypes has important clinical ramifications, as they have variable treatment options and different prognoses. For example, patients with Luminal or HER2 subtypes have options for targeted and hormonal therapies, while patients with basal subtyped tumours gain no advantage from these treatments and only have chemotherapeutic options.[64] Prognostic signatures, like PAM50 and OncotypeDX, have been developed to give patients and clinicians improved treatment options.[64]

1.7 Publicly Available Gene Expression Datasets

Newer and larger sources of publicly available transcriptomic data have enabled breakthroughs in "Precision Medicine" and will be crucial to the future improvements of the treatment and prevention of BC.[65] The predominant sources of these datasets are, by large, initiatives and national projects including the Cancer Genome Atlas (TCGA) and the International Cancer Genome Consortium (ICGC). Large scale efforts to collect and make available databases will facilitate new waves of *in silico* hypothesis testing for researchers without the ability to develop their own large-scale cohorts. This data is fundamental to improving the scientific process of individual and personalised medicine, as well as powerfully proving or disproving currently held assumptions.[65]

There have been several recent seminal database publications with far reaching impact and utility. The Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) is one such effort that sequenced 173 genes in 2433 patients, along with expression data, copy number aberrations (CNA), and long-term follow up.[66] This study identified 40 driver mutation genes and correlations between the measured genomic information and survival to assess the importance of genome-based analysis and stratification of breast cancer subtypes for therapeutic purposes.[66] The dataset was made publicly available in 2012, with sequencing data added later, and led to the development of the IntClust (IC10) novel sub-grouping classifier, which was based on upstream driver genes identified by the integration of genomic and transcriptomic data.[67]

The Cancer Genome Atlas collected DNA copy number, methylation information, array and sequence based mRNA expression, micro RNA and protein/phosphoprotein expression information for 825 primary BC patients.[65,68] They identified specific and significant mutations in TP53 (Tumour Protein 53), PIK3CA (phosphatidylinositol-4,5-bisphosphate 3-kinase catalytic subunit alpha) and GATA3 (GATA Binding Protein 3) across all BCs and found evidence for four primary BC subtypes, each distinct and with significant molecular heterogeneity.[68] Comparisons were also drawn between basal-like BC and high serous grade ovarian cancers with regard to molecular commonalities and they hypothesised the opportunity for shared therapeutic options.[68] Additionally, they concluded that clinically observed heterogeneity in BC was mostly within and not between distinct BC subtypes.[68]

Outside of major studies, online repositories exist to host individual datasets from researchers and clinicians including the National Center for Biotechnology (NCBI), the European Bioinformatics Institute and, at a smaller scale, The Swiss Institute of Bioinformatics. These accept deposited data and host datasets of genomic information, and allow for improved access to the available public data. Large scale changes and additions to these repositories can be found in the Nucleic Acids Research (NAR) annual database overview. These additions show the rate of growth of available genomic information and highlight the increase in availability of such data. Figure 1.7.1, Page 12 shows the rate of additions and updates to the NAR over a four year period.[69–72] While these represent different types of datasets (proteomic, transcriptomic, and genomic), the rate at which new datasets are accepted shows the trend in the availability of data. The available amount of information for bioinformatic analysis is growing yearly and is opening new pathways to treatment and therapy. Other large scale dataset projects have taken the open source approach, creating useful, open access tools for generating cancer signatures in BC and in other tumours.[73,74]

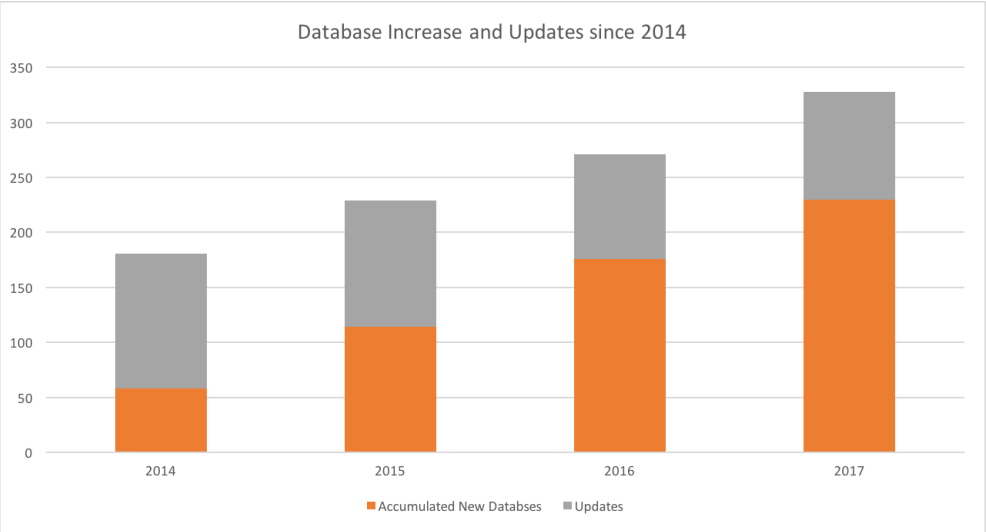


Figure 1.7.1: NAR Year on Year Growth. The 2017 publication of the NAR database issue saw the addition of 54 new databases and 98 updates to existing databases. It saw a steady increase from 2016 (62 new databases and 95 updates) and 2015/2014 (56 and 115, and 58 and 123 new databases and updates, respectively).

1.8 Gene Expression Profiling Methods and Datasets

Gene expression profiling gives a snapshot of a tumour's cellular activity as a function of the level of activity of all measured genes at a specific point in time. This has enabled the detection of differences in gene expression levels among different types of BC and helped to create subtype specific profiles.[75] This technology has helped to better understand the heterogeneity of BC and is offering new clinical tools for the prognosis and treatment of BC.[76] Some profiling techniques, such as OncotypeDX (Genomic Health Inc, Redwood City, CA), MammaPrint (Agendia, Amsterdam, The Netherlands) and PAM50 (Prosigna), are commercially available and showing validated utility in this area, as well as being refined through ongoing clinical trials.[76] This progress is aided by the development of newer and more efficient means of sequencing large scale genomic information.

Microarray platforms for generating the necessary expression level data, like Affymetrix, use thousands of nucleic acid probes to hybridise complementary nucleic acid sequences in solution and measure their relative concentrations to calculate gene expression levels estimated indirectly by observing the hybridisation events.[77] Beadchip sequencing techniques, like Illumina, use thousands of 3 micron silicon beads dispersed in random wells on a testing substrate. These wells are covered in hundreds of thousands of oligonucleotide primers and the

concentration of binding at each bead is measured with fluorescence to estimate the amount of the gene bound.[78] Lastly, RNAseq utilises deep-sequencing to isolate sample RNA, generate complementary cDNA and sequence the fragments, before aligning the reads to map the sample transcriptome. RNAseq has single base resolution and is sensitive enough to distinguish isoforms, but is the most costly of these technologies.[79]

These techniques have all helped to resolve detail in expression level data from patient samples and have made the available predictive and prognostic tools possible.

1.9 Predictive Biomarkers for Patient Response

Biomarkers for cancer fall into three categories; predictive, prognostic or diagnostic.[80] Predictive biomarkers are utilised to predict response to specific therapies like trastuzumab, where the activation of HER2 pathways in BC is indicative of response.[80] Prognostic biomarkers are not treatment specific; instead, they attempt to quantify the risk of future relapse or recurrence of cancer to a patient and help clinicians make informed treatment decisions.[80] Lastly, diagnostic biomarkers are used to ascertain if a patient has a specific form of a disease.[80] As has been touched on, most biomarkers and prognostic tools are developed using single time point patient and tumour information and do not take into account any on-treatment information. The addition of on-treatment information may be advantageous for the development of robust new biomarkers to aid clinicians.

1.10 Prognostic Gene Expression Signatures for Breast Cancer

The most established gene expression based prognostic tools for BC are Oncotype DX and MammaPrint. Oncotype DX is a genomic assay (Genomic Health, Inc., Redwood City, CA, USA) used to predict the likelihood of BC recurrence and aids in the treatment selection process for patients with ER+ disease.[81] Oncotype DX works by comparing the expression of 16 key genes to the expression of five reference genes and algorithmically generates a recurrence score based on these measurements.[82] The effectiveness of Oncotype DX has been independently validated in node-negative patients to accurately predict risk of recurrence regardless of patient age or tumour size.[82] Additionally, the National Surgical Adjuvant Breast and Bowel Project (NSABP) B-20 trial showed that the calculated recur-

rence scores could be used to identify the patients to whom adjuvant chemotherapy conveyed the most advantages over endocrine therapy alone. It found that patients with high risk of relapse gained the most added benefit from adjuvant chemo (RS (relapse score) greater than or equal to 30%).[83] MammaPrint is a microarray-based, recurrence risk test reliant on the expression of 70 genes that make up its specific signature.[84] This test functions as a means of identifying patients with a low or high risk of distant metastasis in order to aid physician treatment selection and minimise unnecessary treatment, as patients labelled as low risk can omit chemotherapy without a reduction in disease-free survival (subsequent fatalities were not caused by a recurrence of the cancer).[85] This result was ratified in the RASTER trial which showed that in the 5-year, distant recurrence-free interval (DRFI), the use of MammaPrint correctly identified low risk patients and the omission of additional chemotherapy did not compromise their outcomes.[86]

While both MammaPrint and OncotypeDX methods suggest patient prognosis and help to identify patients who did not need additional chemotherapy, there are a few key differences between the techniques. Both can now be applied to FFPE (Formalin Fixed Paraffin Embedded) samples, though, initially, MammaPrint was more limited in application, as it could only be performed on fresh frozen samples.[87] Fresh frozen samples are usually "higher quality" samples as they experience less degradation from environmental exposure post-resection, which usually leads to better coverage during subsequent analysis. Additionally, while both are primarily for ER+ tumours, MammaPrint was developed with a younger cohort, and is most appropriately applied to patients under 61 years of age.[87] Finally, trials for both methods are underway to identify if either is appropriate for ER- tumours, which would substantially differentiate between these methods and increase their utility in the future. Other techniques for gene expression profiling that are less well adopted include the Genomic Grade Index and PAM50 (Prosigna).

PAM50 (Prosigna; Nanostring Technologies, Seattle, WA, USA) is another recurrence scoring test for ER+ patients designed from 50 key genes and is appropriate for postmenopausal women undergoing endocrine therapy.[87] PAM50 can report the intrinsic subtype of the tumour[88] and has the added advantage of being performable locally,[87] whereas MammaPrint and Oncotype DX must be performed at specified testing centres. The Genomic Grade Index (MapQuant Dx, Ipsogen, France) represents an entirely different strategy, which is currently used to better define histological grade assessments.[89] Instead of assigning grades to tumours, it defines low and high molecular grade risk groups based on a 97

gene signature (or abbreviated 6 genes RT-PCR signature) in order to separate tumours into more distinct prognosis groups.[87,89]

EndoPredict is a validated assay tool used in the prediction of metastasis for patients with ER+, HER2- BC.[90] EndoPredict incorporates both clinical and genomic information to calculate the risk of metastasis on a per patient basis.[90] This information can be used to inform treatment options and has already been used to advise treatment progression on patients in the study cohort, resulting in additional chemotherapy for some high risk patients or endocrine-only treatment plans for lower risk individuals.[90] Another predictive tool is the Breast Cancer Index (BCI) which is a gene expression based signature for predicting the early and late lifetime disease recurrence of BC in ER+ patients.[91] This test was shown to perform well from the time of diagnosis for the prediction of high and low risk disease recurrence.[91] This information can be used to help inform treatment decisions as well as the suitability of long term adjuvant endocrine therapy.[91] Newer tests are being developed for specific treatment cohorts that have refined applicability and improved resolution of patient prognosis, which can potentially provide significant improvements in the treatment decision making process. One such test is the OncoMasTR, which used network delineation to identify the underpinning genes that are the upstream drivers of BC in a cohort of ER+ LN- patients.[92] The seven genes derived from this transcriptional regression were shown to be more sensitive when predicting the need for additional chemotherapy for women with no lymph node involvement.[92] OncoMasTR can be used in the future to assist in treatment decision making for similar patients in an effort to avoid unnecessary chemotherapy.[92]

1.11 Sequential Sampling

Sequential sampling, taking several samples over time to monitor the change in tumour biology, is offering new insights into the effects of treatment on BC, see Figure 1.11.1 on Page 16 for an illustration of the approach. Initially uncommon, the collection of multiple biopsies from the same patient has become more routine with the rise in neoadjuvant therapies.[93] Studies of neoadjuvant therapies and pre-surgical treatments allow for unique *in vivo* analysis of tumour treatment response,[94] as well as the possibility of predicting the response to treatment earlier in the treatment cycle.[95]

Neoadjuvant studies of sequential samples take place after diagnosis but before surgery, providing a “window of opportunity” that offers the possibility

of observing translational changes that are solely the response of the tumour to treatment.[96] Measuring these changes can potentially indicate if a tumour is likely to respond to treatment or generate resistance and both outcomes can help inform patients and clinicians on how to proceed with treatment.[97–99] Though the exact structure of sequential studies will differ, diagnostic biopsy samples as well as one or more on-treatment (neoadjuvant samples taken after the initial treatment and before total resection) samples and the surgical/resection sample are usually collected. Acquisition of these samples can prove difficult for patient retention in trials, especially in the CT (chemotherapy) setting, where treatment side effects are much more pronounced and additional patient stress is a contributing factor.

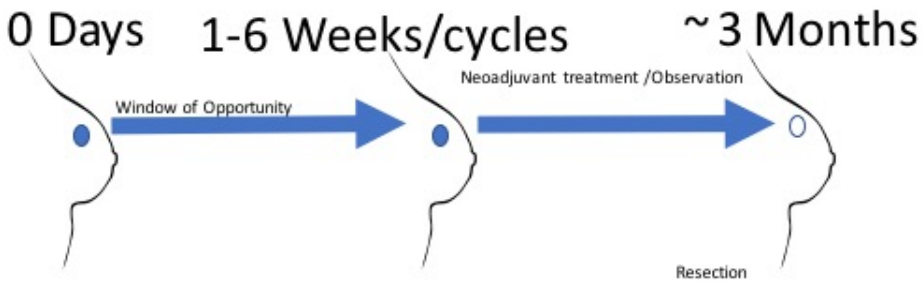


Figure 1.11.1: Window of Opportunity Schematic for Neoadjuvant Therapy. The general structure of neoadjuvant trials and treatment strategies are illustrated here, with a pretreatment opportunity to sample the tumour positive tissue and measure the effects of therapeutics prior to surgery.

1.12 Opportunities and Challenges of Patient-Matched Samples

Sequential samples offer increased resolution for the change in tumour biology with treatment. Most studies focus solely on the presentation of the tumour at diagnosis.[95] However, by looking at the on-treatment changes, important expression pathways that are linked to different clinical outcomes can be observed.[100] Additionally, these sequential samples can offer insight into the molecular changes that drive tumour evolution in patients. Specifically, Guttierez et al. [101] recently reported significant HER2 and MAPK signalling in patients with acquired resistance to tamoxifen in matched pairs of breast samples.

Sequential sampling offers a unique opportunity to study drug efficacy, predict response and provides an increased rate of predictive biomarker

generation.[96] The latter was demonstrated in BC by researchers at The University of Edinburgh through the addition of on-treatment expression information to enhance predictive power.[102] This resulted in a four-gene biomarker that is 91% accurate in validation and also significantly improved prediction of both recurrence and breast cancer-specific survival rates.[102] Further studies within this field are required to establish if this only holds true for ER+ cohorts or if this is more generally applicable. Challenges involved in collecting sequential samples are varied; additional sampling is inherently more expensive and patients are subjected to additional stress. Early on-treatment biomarkers for response would reduce the incidence of inappropriate treatment selection and remove the burden of multiple biopsies. Furthermore, Turnbull et al. have shown that a two-protein signature can be highly effective for predicting response in patients at early time points in BC therapy.[103]

1.13 Availability of Datasets

Studies of sequential time points like those discussed previously are rare compared to single time point studies. Table 1.1 on Page 19 highlights some of the available datasets of sequentially sampled patient studies.[96,102,104–112] It is worth noting the discrepancy between endocrine and chemotherapy in terms of patient cohort size. Endocrine studies have a higher success rate for sequential biopsies because of the reduced levels of stress experienced by patients compared to chemotherapy. Endocrine therapy patients are also at a reduced risk of post biopsy complications and secondary immunosuppression. Practically, this means more chemotherapy based trials of sequential samples must be aggregated in order to reach similar levels of significance for any subsequent results, assuming a significant effect size. Not all datasets are annotated well; some lack phenotypic data for each patient and others defined response status (if available) in different ways, e.g. radiological vs pathological complete response. Therefore, validation across these disparate datasets becomes more challenging as common end points need to be defined to draw direct comparisons. Samples are also processed across different platforms, leading to the inclusion of batch effects that need to be considered and removed[113] in order to adjust for technical variance while retaining the underlying biological information.[114] Lastly, while sequential studies all follow general paradigm of pre-to-on-treatment, time points can differ between studies due to different study designs and sampling requirements. All studies have a pretreatment biopsy sample but the first on-treatment sample is not always uniform nor are subsequent time points. Furthermore, datasets may contain a variety of different on-treatment time point samples due

to different trial and study requirements. Finding a lowest common denominator of treatment samples is necessary for direct comparisons, but will inevitably reduce the number of available samples.

Treatment Type	Days on Treatment	Number of samples	Platform	Date and Study
Letrozole	0,14,90	89,89,89	Affymetrix U133A and Illumina HT-12	Turnbull et al. 2015 (Miller et al. 2009)
Anastrazole	0,14,90	112,97,29	Illumina WG-6	Dunbier et al. 2013
Anastrazole	0,9-63	26,26	Illumina DASL	Morrogh et al. 2012
Celecoxib	0,14-21	22,22	Affymetrix U133 Plus 2.0	Brandao et al. 2013
Randomised Letrozole vs. Anastrozole	0,14	34,34	Custom 17K	Mackay et al. 2007
Everolimus	0,14	23,21	Illumina Ref-8v2	Sabine et al. 2010
Docetaxal and capecitabine	0,21	21,14	Affymetrix HG-U113 Plus 2.0	Korde et al. 2010
Epirubicin or cyclophosamide	0,84	32,25	Agilent 44K	Stickeler et al. 2011
Doxorubicin/cyclophosamide or doxorubicin/docetaxal	0,126	31,17	18K NKI microarrays	Hannemann et al. 2005
No-Intervening Treatment	0,13-53	37,37	Illumina HT-12 v3/4	Pearce et al. 2016
Anthracyclines and Taxanes	0,14,42,90	34,12,23,24	Ampliseq	Bownes et al 2019
Anthracyclines followed by Taxanes	0,1-4,90	221,36,32	Custom cDNA	I-SPY trial
Anastrazole, Letrozole, Exemestane	0,14,28	58,58,60	Agilent Array	Ellis et al. 2017

Table 1.1: Descriptive Table of Some Available Sequential Breast Cancer Datasets. Many of the available datasets for matched pre- and on-treatment primary BC are quite small, meaning the individual results of analysis of these data are potentially lacking in power or significance. The Stickeler, Mackay and Hannemann studies were all performed on FFPE samples, the rest are sourced from fresh frozen cohorts.

1.14 Dataset Integration

A significant problem for sequentially sampled dataset analysis is the size and availability of data. There are multiple obstacles that prevent a simple integration of data to improve sample size; primarily, batch effects and the influence of different processing platforms or methods. However, when discussing the impact on time-dependent data, on-treatment effects become a problem. Contemporary work ongoing at the University of Edinburgh is showing that integration of compositionally diverse datasets can skew subtyping results.[115] Additionally, integrating treated and untreated data together can reduce distinctions between the sample populations. All of these factors will have to be combated, but with successful integration of sequential datasets, the ability to gauge the effect of treatment and leverage the pairwise patient response for risk stratification is improved.

Dataset integration is an attractive target for furthering bioinformatics, as it is desirable to create larger, higher value datasets. These datasets can improve the statistical significance of results and enhance detection of low frequency anomalies from heterogeneous data. However, integration can also introduce bias and statistical artifacts due to the nature of batch correction methodologies seeking to normalise parametric distributions. To better illustrate this point; a two dimensional representation of multiple expression sets would appear as multiple distinct clusters of points. Differential analysis on this unintegrated data would find the differences between the clusters, not the differences between labelled samples across all of the datasets. Additionally, there can be large cell-type heterogeneity and sampling impurities, which, when not considered, can introduce additional noise as these samples are essentially “mislabelled”. Such samples are clearly identified through visualisations of the data and the effects of integration through techniques like principal component analysis. Small and disparate datasets are a significant bottleneck to generating novel biomarkers and prognostic signatures.[116] As no standard for processing platforms exists, the need for a robust integration method becomes more evident as more high quality genomic data is made available.[117] As expression data can vary greatly between platforms in terms of scale and structure, integrating, comparing and validating results across independent studies remains challenging.[117] Batch effects can confound the integration of disparate datasets and removing systematic differences between datasets is necessary for the robust integration of data. As demonstrated by Sims et al., removing batch effects and reconciling systematic biases between expression sets allows for the direct integration of raw expression data with improved statistical significance for downstream

analysis.[95] Examples of these batch creating effects include platform or institute of the data origin, sampling errors, multiple biopsy variation, cell type heterogeneity/composition. These can all have profound effects of the resultant gene expression levels, and especially in comparative analysis.

1.15 Dataset Integration Improves Statistical Significance

Dataset integration allows for the creation of larger and more representative cohorts for statistical analysis and would eliminate the effects of overfitting on a resultant biomarker, due to the wide nature of gene expression data. The small size and specificity of many modern datasets give rise to incomplete clinical tools that are difficult to validate in exogenous data.[118,119] Currently, several prognostic and profiling signatures have been shown to have some level of concordance in independent datasets.[118] However, it has been established that much larger sample sets are necessary to make true consensus biomarkers.[120] Additionally, many prognostic markers and profiling signatures do not validate in other datasets, even when the differences between the sets are minimised.[121] This is frequently due to overtraining of prognostic markers on small datasets that inhibit their general utility in disparate/dissimilar datasets.[121]

There are a number of technical difficulties that plague genomic data integration. Due to differences in pre-processing pipelines and gene- and probe-level annotation, feature size of integrated data is often diminished in order to retain complete feature coverage.[122] Several groups have reported that integration of expression microarray data has positive effects on statistical power, the ability to identify differentially expressed genes of interest and more robust, reproducible results.[122–124] While there is no standard for the method of integration, it has been shown that, as long as the integration methodology is rational and carefully executed, the improvements to downstream analysis are significant.[122]

There are few existing examples for the integration of sequential samples. Primarily, this may be due to the relative scarcity of sequential datasets and the added confounding factor of patient and treatment response heterogeneity. Integration must be carefully considered, as it has been shown previously that dataset composition can affect the accuracy of expression derived subtypes in BC.[125] As ongoing treatment changes the predicted intrinsic subtype, integration of different time points must be handled carefully to avoid normalisations across treatments that could potentially distort gene expression measurements and in-

troduce, rather than remove, bias.[125]

1.16 Using On-Treatment Information May Enhance Prediction of Response or Prognosis

There is a substantial precedent for the clinical use of gene expression profiling. This is primarily based upon correlated relationships between pretreatment patient samples, post treatment patient samples, and long-term survival analysis. While this methodology has proven suitable in the generation of successful expression signatures for prognostic prediction (Onxotype DX, MammaPrint etc.), the scarcity of on-treatment information has made the generation of signatures from sequentially sampled data significantly more rare. Developing early on-treatment profiling signatures for response in both neoadjuvantly administered chemotherapy and endocrine therapy could improve patient care and reduce over prescription. As BC is a multifaceted disease, a single biomarker is unlikely to be successful at predicting response in a molecularly diverse setting. Fortunately, there is now access to sequentially sampled data for both treatments types, which facilitates the generation of niche markers for both treatment paradigms.

A new landmark study in this field, the POETIC trial, is a phase III, multicentre, randomised trial of ER/PR+ BC measuring the effect of a perioperative aromatase inhibitor with on-treatment changes to KI-67.[126–128] This study builds on the established results of the IMPACT trial, which suggests that on-treatment changes in KI-67 at two weeks are strongly suggestive of patient outcome.[129] Ellis et al. have also demonstrated that on-treatment information can be informative for the accurate prediction of response to endocrine therapy.[130] In a 2017 study, they examined the expression levels of KI-67 in patients at two and four weeks and found that patients with elevated KI-67 levels (higher than 10%) were exhibiting endocrine resistance and were triaged to neoadjuvant chemotherapy.[130] The triaged cohort experienced lower rates of pCR (5.7%, 2/35 patients).[130] Individually, these studies support the use of on-treatment gene expression monitoring for improved patient stratification.

Studies of integrated pre- and on-treatment information are still in their infancy. However, Turnbull et al. revealed that on-treatment proliferation markers could be combined with patient-matched pretreatment markers to accurately predict the clinical response and recurrence-free survival of patients.[102] In this study, the tests utilising on-treatment information outper-

formed pretreatment-only diagnostic tools, in both accuracy and specificity.[102] The on-treatment signature developed was found to be 96% accurate, while the pretreatment-only signature was 93% accurate in a local dataset.[102] However, in validation, their four-gene, pre- and on- treatment signature outperformed a similar pretreatment only signature by 14%.[102] Continuing work by Sims et al. has combined NPI prognostic groups with their two-protein signature to further refine the predictive capabilities of these on-treatment signatures.[103] Figure 1.16.1 on Page 23 highlights the overall structure of this study and how to follow the results that are presented.

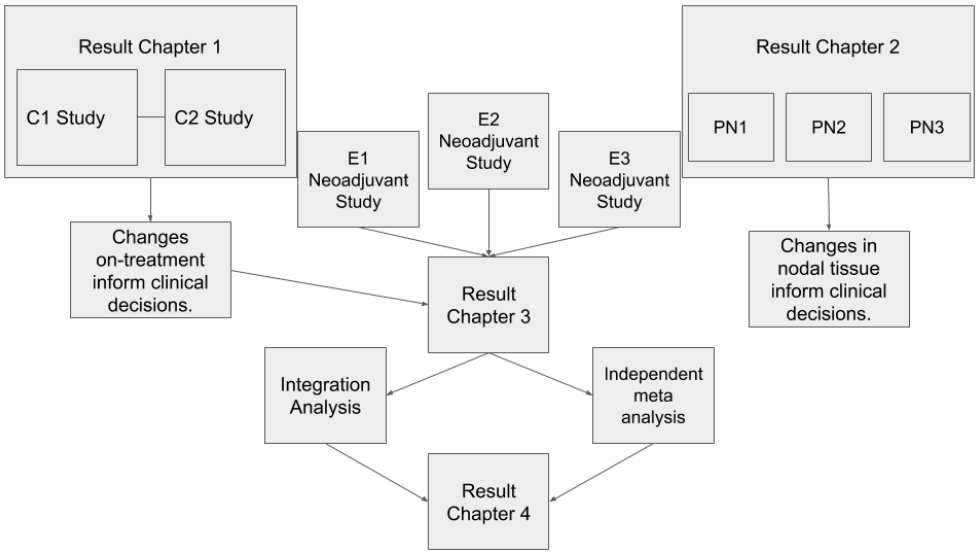


Figure 1.16.1: Flow Diagram of the Overall Structure of Thesis. This thesis contains four results chapters which correspond to Chapters 2-5. Chapter 2 feeds directly into chapter 4, which subsequently feeds into chapter 5. Results chapter 3 is independent in regards to the data analysed, but has similar themes of comparing the value of matched non-primary tissue samples for prognostic purposes.

1.17 Thesis Hypothesis, Aims and Outcomes

This thesis aims to assess the value of additional on-treatment sampling of neoadjuvantly-treated BC patients for the purpose of biomarker identification and risk stratification to prove the following thesis:

> *If patient-matched on-treatment or lymph node positive samples are informative for the classification and characterization of breast cancer, then they should facilitate improved differential and statistical analysis of breast cancer.*

Sequential sampling of neoadjuvantly treated breast cancer should capture more informative transcriptional changes, which will be more indicative of tumour response to therapy. This will enable improved identification of patient response and enhance clinical decision making options, furthering the goal of personalised medicine. This project aims to enumerate the value of multiple sampling in matched primary and lymph node biopsies to provide additional prognostic risk information. Lastly, in an attempt to improve future analysis of BC, this thesis aims to exhaustively study the possibility of integrative analysis and/or comparative meta-analysis to improve the resolution of treatment specific changes and more accurately model and predict risk to patients.

2 | On-treatment Biomarkers can Improve Prediction of Response to Neoadjuvant Chemotherapy in Breast Cancer

2.1 Abstract

Background

Neoadjuvant chemotherapy is increasingly given preoperatively to shrink breast tumours prior to surgery. This approach also provides the opportunity to study the molecular changes associated with treatment and evaluate whether on-treatment sequential samples can improve response and outcome predictions over diagnostic or excision samples alone.

Methods

This study included a total of 97 samples from a cohort of 50 women (aged 29–76, with 46% ER+ and 20% HER2+ tumours) with primary operable breast cancer who had been treated with neoadjuvant chemotherapy. Biopsies were taken at diagnosis, at 2 weeks on-treatment, mid-chemotherapy, and at resection. Fresh frozen samples were sequenced with Ion AmpliSeq Transcriptome yielding expression values for 12,635 genes. Differential expression analysis was performed across 16 patients with a complete pathological response (pCR) and 34 non-pCR patients, and over treatment time to identify significantly differentially expressed genes, pathways, and markers indicative of response status. Prediction accuracy was compared with estimations of established gene signatures, for this dataset and validated using data from the I-SPY 1 trial.

Results

Although changes upon treatment are largely similar between the two cohorts, very few genes were found to be consistently different between responders and non-responders, making the prediction of response difficult. AAGAB was identified as a novel potential on-treatment biomarker for pathological complete response, with an accuracy of 100% in the NEO training dataset and 78% accuracy in the I-SPY 1 testing dataset. AAGAB levels on-treatment were also significantly predictive of outcome ($p = 0.048$, $p = 0.0036$) in both cohorts. This single gene on-treatment biomarker had greater predictive accuracy than established prognostic tests, Mammaprint and PAM50 risk of recurrence score, although interestingly, both of these latter tests performed better in the on-treatment rather than the accepted pre-treatment setting.

Conclusion

Changes in gene expression measured in sequential samples from breast cancer patients receiving neoadjuvant chemotherapy resulted in the identification of a potentially novel on-treatment biomarker and suggest that established prognostic tests may have greater prediction accuracy on than before treatment. These results support the potential use and further evaluation of on-treatment testing in breast cancer to improve the accuracy of tumour response prediction.

Overview

A generalised diagram of for this work, especially with regards to the generation of the most vital outcomes is presented in Fig. 2.1.1, on Page 27. This high level over view shows the flow of the paired samples from patient fold change values through to candidate gene lists for response and lastly validation testing in new testing data.

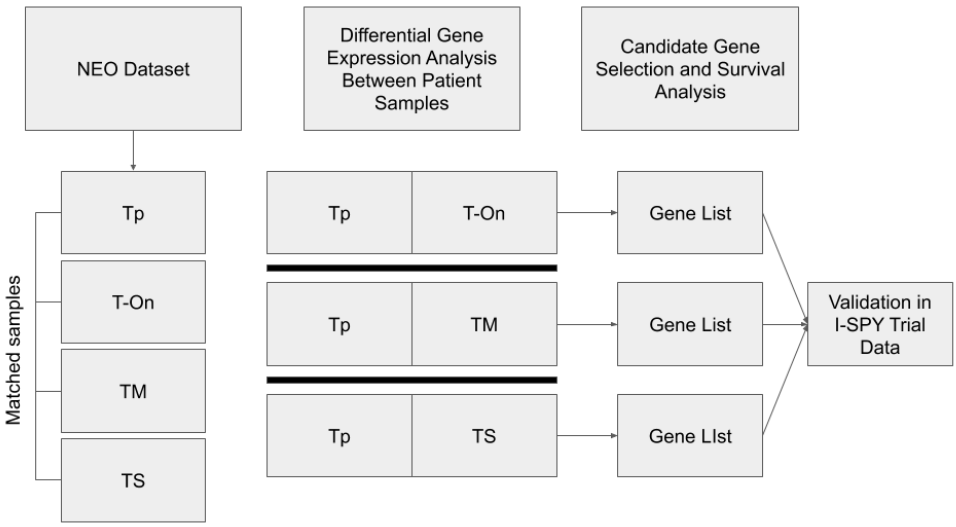


Figure 2.1.1: Overall Study Design and Workflow. This workflow diagram illustrates a high level overview of the analysis of the NEO trial data. Starting on the left, there are the matched patient samples across time, Pretreatment (TP), Early On-Treatment (T-On), Mid-Treatment (TM) or the surgically resected sample (TS), then differential gene expression analysis of the time point pairs, followed by candidate gene discovery and risk modelling on the output gene lists. Lastly, these results are validated in a contemporary study, the I-SPY trial.

2.2 Background

Chemotherapy is among the most common effective treatments for breast cancer, alongside radiotherapy, hormone therapy, and targeted treatments. Neoadjuvant chemotherapy (NACT) is given prior to surgery to reduce the tumour burden and provide early information on treatment response.[131] Studies have shown that patients with tumours that have a pathological complete response (pCR) following neoadjuvant chemotherapy are much less likely to recur than those in women with residual disease.[132] Neoadjuvant chemotherapy is now considered as the standard of care in breast cancer and has seen a rise in recent years with data from powered studies suggesting that the pathological complete response achieved following neoadjuvant chemotherapy might be a surrogate of good prognosis.[133] Recent meta-analysis has shown significant tumour response and an increase in the rate of breast-conserving surgery following NACT with good rates of long-term local recurrence (5.5% vs. 15.9% adjuvant chemotherapy), but an increase in the rate of short-term local relapses (1.35 RR 0–4 years, 1.53 RR 5–9 years).[134]

Neoadjuvant treatment provides a “window of opportunity” (Fig. 2.4.1, A, on

Page 35), where sequential sampling of a tumour enables observation of the changes that occur in response to treatment to be measured and considered in the context of response and outcome.[135] Neoadjuvant therapy studies and pre-surgical treatments allow for a unique *in vivo* analysis of tumour treatment response,[136] as well as the possibility of predicting the response to treatment earlier in the treatment.[135] It has been suggested that on-treatment biomarkers may be superior to those measured before exposure to treatment.[133,137] On-treatment information has already been shown to be informative for the accurate prediction of response to endocrine therapy.[138] Here, it was found that patients with elevated KI-67 levels (higher than 10%) at two or four weeks exhibited resistance to endocrine therapy and were triaged to neoadjuvant chemotherapy.[138] It has also been demonstrated that there is potential of on-treatment biomarkers by developing a four-gene signature which combined pre-treatment expression levels or two biomarkers (IL6ST, Interleukin 6 Signal Transducer, and NGFRAP1, Nerve Growth Factor Receptor Associated Protein 1) with patient-matched 2-week on-treatment expression levels of two proliferation markers (ASPM, abnormal spindle microtubule assembly, MCM4, Minichromosome Maintenance Complex Component 4) to accurately predict the response to endocrine therapy in a blinded independent validation set.[137]

To date, gene expression-based studies of neoadjuvant chemotherapy treatment have largely been limited to studying the association of pre-treatment samples with pathological response.[139,140] Patient-matched sequential sampling gene expression studies have been previously attempted; however, they have not evaluated the predictive capacity or proposed new on-treatment predictive biomarkers.[141–143]

In this study, we present the largest sequentially sampled patient-matched analysis of neoadjuvant chemotherapy-treated breast cancer tumours to evaluate whether on-treatment biomarkers can improve the accuracy of predicting response before resection. Numbers of patients with sequential breast tumour samples are limited, but we compare and validate our results with the data from the I-SPY 1 trial.

2.3 Materials and Methods

2.3.1 Patients, Response Criteria, and Samples

The NEO study consists of 50 breast cancer patients with sequentially sampled biopsies at four time points, pre-treatment (TP, 34 samples), 2 weeks on treat-

ment (T-On, 12 samples), mid-chemo (TM, 23 samples), and at surgical resection (TS, 24 samples). There were three clinically defined response statuses: complete responders (pCR by resection), good responders (tumour volume reduction, but lack of pCR), and non-responders (progressive disease or small tumour volume changes on treatment, these included patients with stable non responding disease). There were 11 TP-T-On, 23 TP-TM and 36 TP-TS pairs. Patients were of mixed histological grade and HER2 status; ages ranged from 29 to 76. Patients were primarily treated with 3 cycles of FEC 100 (5 Fluorouracil, Epirubicin, Cyclophosamide) and docetaxel 100 with Herceptin, where appropriate. Three patients received paclitaxel, one patient received additional carboplatin, one patient received Epi-cyclophosphamide and paclitaxel, and one patient received docetaxel and cyclophosphamide. Eligible patients were women with histologically confirmed invasive breast tumours and with no evidence of distant metastatic disease, no prior history of malignancy, and fit enough to receive chemotherapy in the opinion of the responsible clinician irrespective of age. All cases were discussed at the breast Multi-Disciplinary-Meeting in Edinburgh Breast Unit at the Western General Hospital, and consensus from this meeting was to be treated with neoadjuvant chemotherapy. Dr Oikonomidou is responsible for the conceptualization and the set up of the NEO study as well as obtaining relevant approvals for the study R&D and Ethics. Dr Oikonomidou and her team identified eligible patients. These patients were consented to the study prior to any sample collection taking place. All samples were collected and processed by Dr Oikonomidou's group. The wet lab experiments were performed by Dr. Carlos Martinez-Perez and Dr. Arran Turnbull. Analysis was performed by Richard Bownes.

Core needle (16-gauge) biopsies were taken from the primary breast tumours before treatment (TP) and between 10 and 14 days after the first dose (T-On) of chemotherapy. A third sample was taken at the mid-chemotherapy point day 20–21 (TM), and finally, a core biopsy was taken from the excision specimen (TS) after it has been removed prior to submission to pathology. Fixed and frozen samples of normal and tumour tissue were collected from all specimens and fresh frozen samples were used for subsequent analysis.

The I-SPY 1 trial is composed of patients with invasive breast cancer > 3 cm, or at least one tumour-positive axillary lymph node.[141] Patients were treated with an anthracycline-based chemotherapy followed by taxanes.[141] Samples were normalised and corrected for background red/green signal; Bioconductor R packages marray and Limma[144] were used for this analysis. From the original 221 patients, only 36 had matching pre- and on-treatment samples, and 39

had matching biopsy and excision samples; pathological complete response was used for response criteria. Pairwise gene expression was handled with SAM (Siggenes) and follow-up analysis with Ingenuity Pathway Analysis from QIAGEN Bioinformatics. I-SPY 1 trial data is hosted at NCBI GEO under accession GSE32603.[141]

2.3.2 Gene Expression Profiling

RNA extraction was performed via Ribo0-RNAseq, and whole transcriptome sequencing was performed with Life Sciences Ion AmpliSeq™ Transcriptome Human Gene Expression Kit. This generated greater than 8 M reads per sample with an average of more than 90% valid reads for 12,365 targeted genes. Most analyses were performed in R[145] using packages available through CRAN and Bioconductor.[146] Outside of the R environment, the stand-alone application Multiple Experiment Viewer[147] was utilised for pairwise ranked product feature selection, and DAVID[148] was used for pathway identification and Gene Set Enrichment Analysis (GSEA). Additionally, the python package scikit-learn[149] was used for unsupervised clustering analysis. Ninety-seven samples were analysed over 13 AmpliSeq chips, but no systematic batch effects were evident and no batch correction was performed within the training data. Gene expression data for the NEO study has been made publicly available at the NCBI GEO data repository under accession GSE122630.

2.3.3 Statistical Analysis Methods

Principal component analysis (PCA) was performed on unfiltered gene lists to reduce dimensionality and visualise differences in response at all times and to identify present differences between patient treatment statuses. Local Fisher discriminant analysis (LFDA)[150] was used at each time point to determine if the response groups could be distinguished with treatment time with a semi-supervised clustering approach, concurrently with class advised K-means clustering. LFDA is a form of supervised dimensionality reduction that maximises between-class scattering and minimises within class scatter, and is a refined version of normal Fisher discriminant analysis;[150] this exploratory analysis was used in order to visualise comparative differences in treatment time, not as a means of feature selection. Pairwise significance analysis of microarrays[151] using the siggenes package in R was used to consider the consistency of differentially expressed genes due to treatment in the sequential patient-matched samples.

The primary method of reporting significance between two groups, or measuring the significance of the difference present between two effect populations in this study, is through a combination of parametric and non-parametric tests. When measuring the effect between groups of unpaired samples, standard T-tests were performed. The five basic assumptions were always tested: The values are on a continuous scale (expression values), the samples are to the best of my knowledge representative of the population as a whole as the only constraint on the sampling process was agreeing to be part of the study and no filter was made for age or ethnicity. Graphically the data conforms to a normal bell shaped distribution after processing, but is slightly long tailed prior to normalisation. The fourth criteria of a reasonably large sample size is satisfied as the other assumptions hold and there are sufficient samples to calculate the test statistic. Lastly, standard deviations on both side of the mean are approximately even suggesting homogeneity of variance. Non-parametrically, a Wilcoxon T-test was used when measuring the differences in expression of paired samples. Assumptions here were easier to satisfy as every sample is inherently paired and come from the same population, additionally, the comparisons are from within pair differences.

Rank Product analysis through MEV was used to identify differentially expressed genes between response classes at each time point. MEV was chosen in this instance over comparable analysis using siggenes due to the level of consistency it provided over contemporary methods with regards to ease of use and granularity to select appropriate comparisons from the data annotation. For the remainder of this work the lists obtained from the MEV-derived rank products implementation will be used for analysis. Successive levels of standard p value (0.05, 0.01, 0.001), without correction for multiple testing, were used in order to determine the number of differentially expressed genes, and at lower p values which time points had the most strongly differentiating genes. Later validation with multiple corrections identified no false positives and thus no results were eliminated. Significance analysis of microarrays was also performed using varying false discovery rates (1%, 5%, 10%) to identify common differentially expressed genes between responders and non-responders across both datasets at each time point in order to perform a “sanity check” on the gene lists derived from a more reproducible source. However, these lists were not included in future analysis. Gene score enrichment analysis was used to validate the time point selection by looking for the highest number of enriched pathways. The gene list from the most differential time point (TM) using the NEO dataset was extracted and used in a random forest model (10,000 trees, m-try as the square root of the feature number) using pCR status as the class label (clinician-identified

pCR and non-pCR). The most deterministic genes for class prediction were fed into a classification and regression tree in order to produce a maximally reduced and repeatable model; this methodology is further described by Turnbull et al.[137] The method employed in Turnbull and co-workers study identified differentially expressed genes between different time point patient sample pairs as a means of dimensionality reduction then used the reduced feature space as the input to a random forest classification model trained on response. The ranked variable importance of each feature is ordered and all genes with a calculated importance to classification over 5% is then fed into a CART (Classification and Regression Tree) model to create a decision tree with maximal pruning for risk stratification on minimal genes.[137] The CART decision tree method was applied to the NEO dataset for training and tested in the independent I-SPY 1 dataset using the same cut-points determined by mean-centring the datasets, this was done make the range of values comparable. This protocol was repeated using the gene list from the pre-treatment-only samples, using the same p values and tree configurations for selection. Survival analysis was performed at different time points using the log-rank test which tests the null hypothesis that the probability of an event is equal in both populations. The log-rank score is calculated as:

$$X^2(logrank) = \frac{(O_1-E_1)^2}{E_1} + \frac{(O_2-E_2)^2}{E_2}$$

In the above equation E represents the expected number of events, and O the observed. Log-rank was selected as this was a standard comparison between two classes,[152] accuracy is presented as balanced F1 scores.

Risk of relapse scores, where specific gene markers are used to estimate the risk of a patient experiencing a relapse, such as MammaPrint, and intrinsic subtypes, such as PAM50, were estimated from the gene expression data using the geneFu R package.[153] Subtypes and risk scores for IC10,[154–156] Pam50,[156] scmod1[157] and scmod2,[158] ssp2003[154] and ssp2006,[155] and the continual and categorical risk values for MammaPrint[159] and rorS[156] were calculated with geneFu. These operate by assigning centroid mappings for each subtype or risk group, transforming the numerical expression matrix data into a reduced space and mapping onto the centroid map. The least distant subtype is then assigned to the tested sample. These methods take matrices of gene expression data as the input and output output sample subtype or profile assignments per column. This information was used for concordance comparisons as well as Sankey diagrams to show the change over time.

IC10 is an integrative clustering algorithm which identifies ten subtypes of

breast cancer marked by their upstream gene driver signatures which have shown distinct clinical outcomes and was originally derived from examining both genomic and transcriptomic data. Pam50 is a tumour profiling test specifically designed for use on ER+ and HER- breast tissue, but which can be used to profile tumours into five different subtypes with differentiating clinical prognoses, or can be employed as the ROR (Prosigna) risk of relapse score which attempts to calculate the likelihood of tumour metastasis. SCMOD1/2 are Subtype Clustering Models based on a series of parameters described by Desmedt, 2008,[157] and Wirapati, 2008,[158] which separate tumours into groups of ER-/HER2-, HER2+ and ER+/HER2- subtypes for prognostic purposed. SSP2003/6 are classifiers that have shown good concordance with clinically defined histologically derived molecular subtypes, identifying tumours as Her2, Basal, LumA and LumB. These were first described by Sorlie, 2003[154] and Hu, 2006[155] respectively. Lastly, Mammaprint is a genomic test which uses a panel of genes to estimate the risk of tumour recurrence and classifies patients according to their estimated risk.

2.4 Results

2.4.1 Gene Expression Differences Between Responding and Non-responding Breast Cancer Tumours Treated with Chemotherapy are Subtle and Time Dependent

Unsupervised principal component analysis was first used to assess whether sequential patient-matched samples from patients receiving chemotherapy (Fig. 2.4.1, B, on Page 35) would cluster by time point or response status. There was no significant grouping of patients according to sampling time: pre, early, or later after chemotherapy in either the NEO or I-SPY 1 studies (Fig. 2.4.1, B, on Page 35). There were no significant differences between the two cohorts in terms of age, grade hormone receptor, and HER2 status, and the subset of patients with mid-chemo samples was not significantly different from the whole NEO cohort (Table. 2.1, on Page 36). Patient-matched samples enable the pairwise analysis to look for consistent changes in the gene expression during treatment. Pairwise significance analysis of microarray analysis using a 10% false discovery rate (FDR) identified a relatively small proportion of overlapping upregulated (5%) and downregulated (4%) genes between the two studies. However, genes that were increased or decreased in response to treatment in one study were also clearly and consistently increased or decreased in the other study, further suggesting it would be difficult to discriminate responders from non-responders.

Indeed, there was no clustering by response status before or during treatment. These results likely reflect the considerable inter-patient differences being substantially larger and more significant than the subtler commonalities in gene expression of a particular time point or response class of each tumour. More encouragingly, semi-supervised LFDA of each time point revealed significant separation on-treatment that was not apparent in pre-treatment samples, indicating that there are meaningful differences between the classes as early as 2 weeks on-treatment (Fig. 2.4.2, A, on Page 38). Complete responders and non-responsive patients were more clearly separated than partially responding patients. These results suggest that there is a potentially greater predictive value looking at on-treatment than pre-treatment biomarkers.

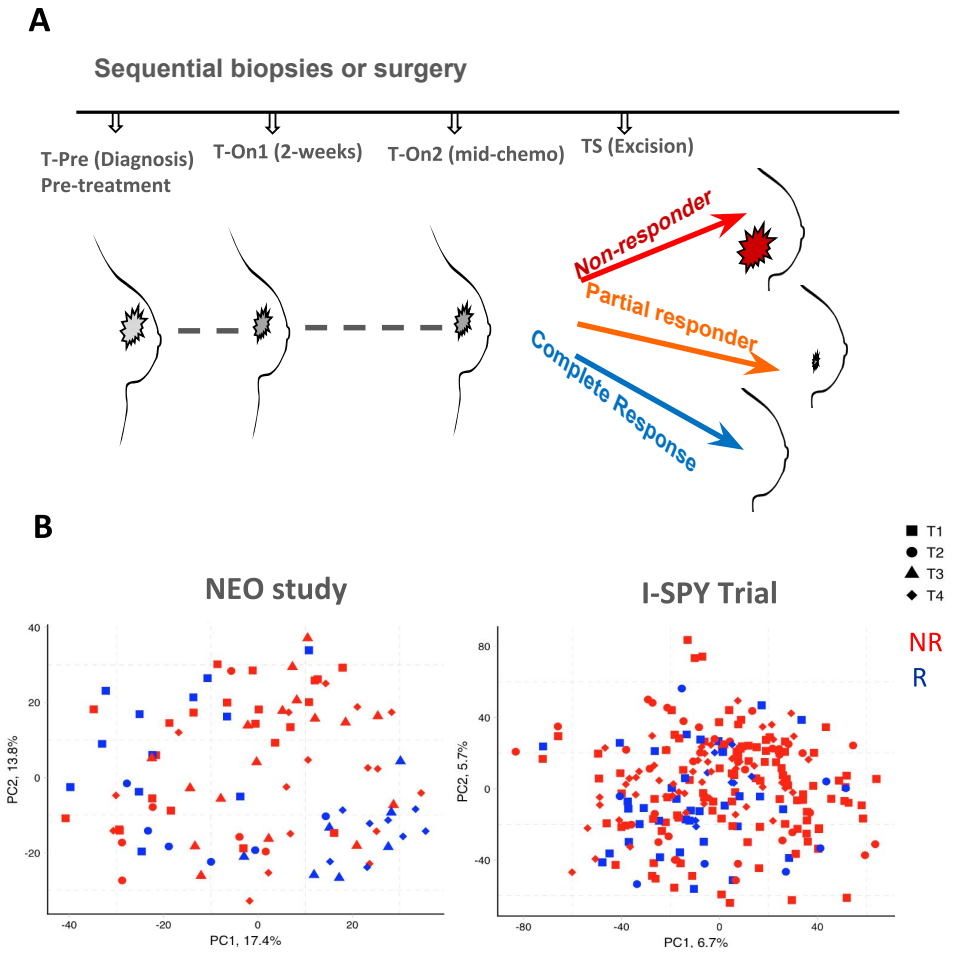


Figure 2.4.1: On-Treatment Study Design and Unsupervised Analysis. A) Neoadjuvant studies all follow similar design scheme, with a pre-treatment sample, potential on-treatment matched samples and a surgical biopsy sample, and with matched patient phenotypic data, where possible. B) Unsupervised clustering by PCA shows no significant trends in response or treatment time.

Characteristics	NEO cohort (50)	NEO TP-TM (23)	*p value*	I-SPY TP-TM (36)	*p values*
Median age at diagnosis	50.8	50.1	0.8	47	
Tumour grade			0.52		0.58
1	0 (0%)	0 (0%)		1 (3%)	
2	22 (44%)	12 (52%)		20 (55%)	
3	28 (56%)	11 (48%		15 (42%)	
Hormone receptor status			0.24		0.66
Positive	23 (46%)	14 (61%)		24 (67%)	
Negative	27 (54%)	9 (39%)		12 (33%)	
HER2 status			0.87		0.64
Positive	10 (20 %)	5 (22%)		6 (17%)	
Negative	40 (80%)	18 (78%)		30 (83%)	

Table 2.1: Summary of Patient Characteristics for the NEO Study and I-SPY Validation Set. There are few significant differences between the characteristics available for the NEO trial and the validation I-SPY data sets. This makes comparisons between the two more logical as there are fewer differences to account for.

2.4.2 Responding and Non-Responding Tumours are More Different Upon Exposure to Chemotherapy

In an attempt to quantify the molecular differences between the response groups at each time point, rank product analysis was performed at different standard p values (0.05, 0.01, and 0.001). This approach was hampered by different numbers of samples at each time point (with T-On having very few samples); however, the number of genes differentially expressed at all p values tended to be greater during rather than before treatment (Fig. 2.4.2, B, on Page 38). Similar results were also seen using 1%, 5%, and 10% FDR (Fig. 2.4.2, B, on Page 38). The biggest differences between the response classes were at TM (mid-chemo), which agrees with the LDFA results, which showed the least amount of overlap of the response classes at TM. Gene set enrichment analysis across the response classes at each time point also demonstrated more enriched pathways after 2 weeks of treatment (29), mid-chemo (30), and resection (29), compared to pre-treatment (18). Next, we sought to examine common differentially expressed genes between responders and non-responders across the two datasets. More genes were significantly differentially expressed (FDR = 10%) between the responders and non-responders on-treatment compared to pre-treatment in the NEO and I-SPY 1 datasets. In accordance with the LFDA results, more significantly differentially expressed genes (1814) were observed between on-treatment samples, with 6% (197), but only one was common between NEO and I-SPY pre-treatment. Examination of the 468 most significantly differentially expressed genes ($p < 0.001$) between responders and non-responders in the NEO dataset at mid-chemo did not clearly distinguish between response groups or time points, further demonstrating that identifying biomarkers of response to chemotherapy is very difficult.

Evaluation of the alteration of the intrinsic subtype assigned to tumours would alter upon treatment. Analysis of the NEO and I-SPY data sets together, it was found that basal tumors were relatively stable with only 2/19 (11%) tumours changing. More tumours were classified as LumA or normal-like on-treatment, which likely reflects a reduction in the expression of proliferation genes during chemotherapy (Fig. 2.4.2, C, on Page 38).

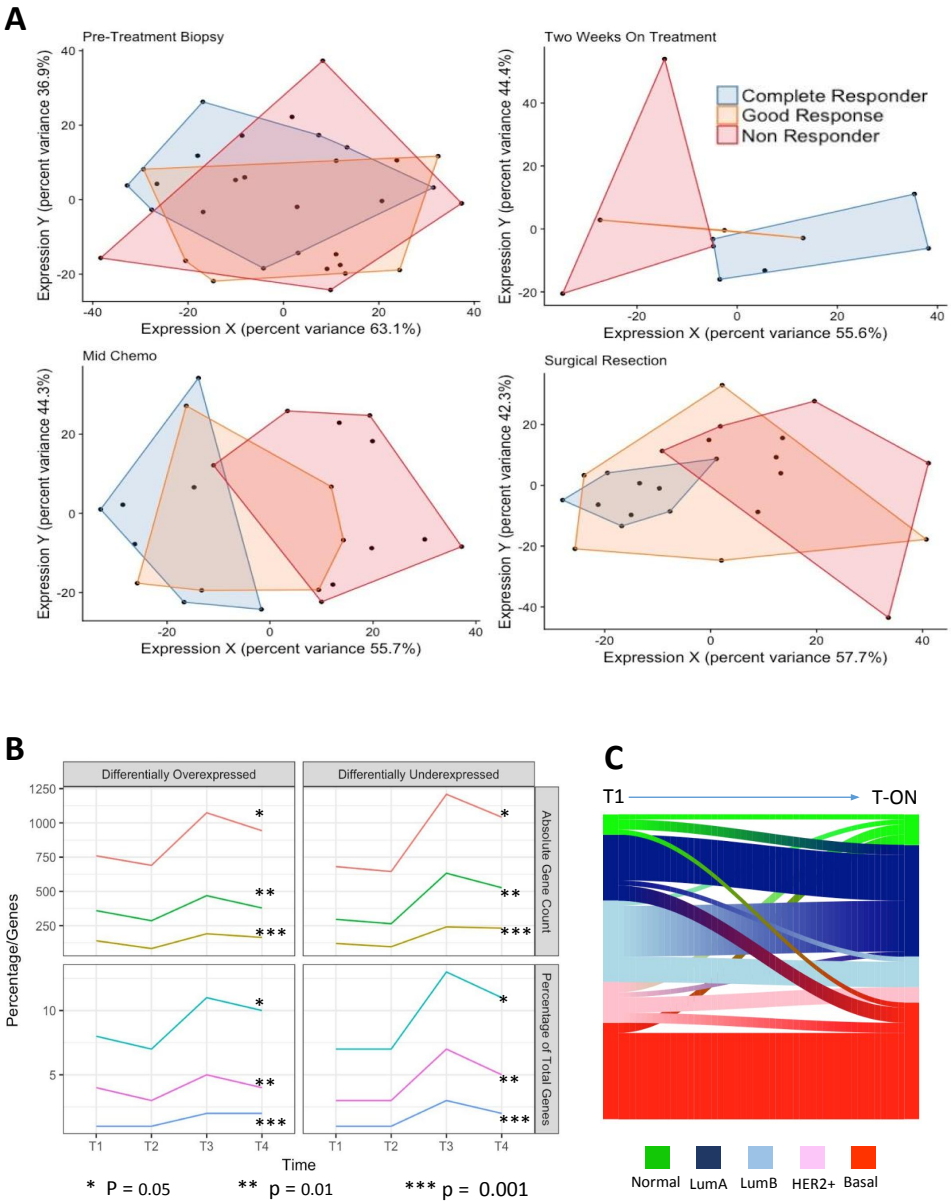


Figure 2.4.2: Initial Evidence of On-Treatment Expression Level Changes. A) LFDA analysis revealed that supervised separation algorithms could significantly identify response specific clusters. B) Differential expression on-treatment was increased for over and under expressed genes compared to pre-treatment. C) This Sankey diagram shows that there is evidence of profile expression changes that can be monitored by the change in calculated molecular subtype.

2.4.3 AAGAB is a Promising Potential Novel On-Treatment Biomarker of Response to Chemotherapy

The mid-chemo gene list from the NEO dataset (1102 genes, p value = 0.01, FDR = 5%) was fed to a random forest model for further feature selection and classification and regression tree (CART) model, which reported AAGAB (Alpha And Gamma Adaptin Binding protein) as the most predictive gene for response prediction in the NEO training dataset with 100% accuracy (Balance F1-Score) for pCR prediction on the mid-chemo samples (Fig. 2.4.3, A, on Page 41). Other methods of classification would have likely been successful, including logistic regression, but the importance of the RF model was in the ability to report easily interpretable importance scores with an intrinsic comparable scale for cut point selection. Validation was conducted independently on publicly available sequentially sampled chemotherapy data from the I-SPY 1 trial,[140] and reported 76% accuracy using AAGAB at the same expression level on the scaled and centred expression data at the on-treatment time point prior to resection (T-On). For comparison, the pre-treatment only sample gene lists were put through the same protocol in order to consider whether highly predictive models could be generated before chemotherapy. IGF1R was the most predictive pre-treatment marker with an accuracy of 74% and 63% in the NEO and I-SPY datasets, respectively (Table 2.2). AAGAB was the sixth most accurate predictor (65%, 57%); receiver operator curves show the relative specificity and sensitivity of this marker pre- and on-treatment (Fig. 2.4.3, A, on Page 41). Gene expression levels of AAGAB were lower in responders across all time points in the NEO cohort but were most significantly different at mid-chemo. In the I-SPY dataset, AAGAB was significantly lower before treatment and at excision (Fig. 2.4.3, C, on Page 41). We wondered whether AAGAB was lower in responders due to a reduction in proliferation, but Pearson correlation analysis with common proliferation-associated genes (TOP2A, BUB1, MKI67, MCM2, FOXM1, and PCNA) demonstrated no significant correlation to any of these genes (Fig. 2.4.3, D, on Page 41), suggesting that AAGAB is independent of proliferation. Survival analysis demonstrated that response status predicted by AAGAB level, at mid chemo in the NEO study and at 2 weeks in the I-SPY 1, was significantly associated with the outcome (NEO $p = 0.048$, I-SPY 1 $p = 0.0036$) (Fig. 2.4.3, D, on Page 41). Interestingly, the level of AAGAB before treatment was not associated with the outcome in either cohort ($p = 0.71$ and $p = 0.2$, Fig. 2.4.3, D). None of the other top ten pre- or on-treatment markers were significantly associated with the outcome in both datasets (Table 2.2); only one gene (ARF5) was associated with the outcome in the NEO dataset ($p = 0.004$). Taken together, the single gene on-treatment biomarker AAGAB appears to outperform novel pre-treatment markers and established prognostic

tests in predicting pCR and long-term outcome to chemotherapy.

	Response Accuracy		Response AUC		Outcome (Log-Rank)	
	NEO	I-SPY	NEO	I-SPY	NEO	I-SPY
Pre-Treatment						
AAGAB	**64.70%**	**56.90%**	0.65	0.58	0.71	0.2
IGF1R	73.50%	62.60%	0.76	0.69	0.36	0.11
CTNNB1	70.60%	49.30%	0.73	0.46	0.6	0.4
SLC20A2	70.60%	55.50%	0.72	0.57	0.063	0.56
HMGCL	67.60%	46.70%	0.67	0.45	0.1	0.97
ST6GALNAC5	67.60%	51.70%	0.69	0.53	0.6	0.28
C1orf51	61.80%	NA	0.61	NA	0.12	NA
KRTCAP3	61.80%	54%	0.63	0.57	0.78	0.78
SETDB2	50%	49%	0.48	0.51	0.29	0.15
FADS2	29.40%	47.80%	0.27	0.5	0.14	0.73
On-Treatment						
AAGAB	**100%**	**78%**	**1**	**0.63**	**0.048**	**0.0036**
ZNF165	87.50%	53.50%	0.91	0.57	0.26	0.7
KRTCAP3	79.20%	52.30%	0.85	0.56	0.81	0.49
RFC2	79.20%	40.10%	0.85	0.35	0.51	0.44
C20orf151	70.10%	NA	0.75	NA	0.36	NA
ARF5	70.10%	43.20%	0.75	0.36	0.0038	0.2
BSPRY	70.10%	47.70%	0.75	0.49	0.47	0.19
NGRN	58.30%	NA	0.66	Na	0.53	Na
CHST7	29.20%	45.50%	0.21	0.52	0.65	0.4
SLC18B1	25%	Na	0.18	NA	0.55	NA

Table 2.2: Relative Performance of AAGAB Pre- and On-Treatment Compared with Other Markers Derived from the Same Methodology. Competitive gene markers for response using the same methods were gathered and compared to AAGAB to gauge performance.

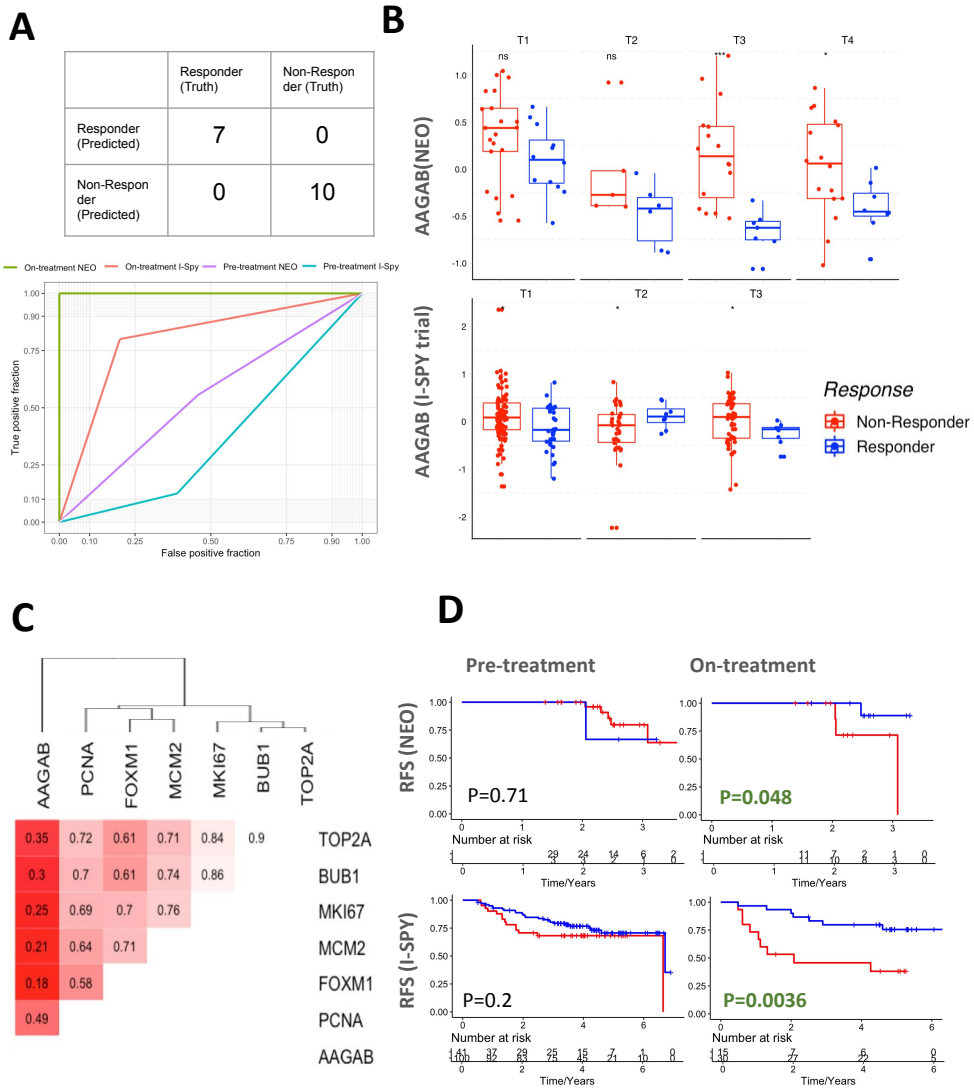


Figure 2.4.3: AAGAB Expression, Response and Outcome Predictive Values. A) AAGAB perfectly separated response categories using the on-treatment expression values as shown in the confusion matrix. Subplot A shows the AUC curves of different separation values. B) The box plots show that there was significant down regulation of the expression of AAGAB in both cohorts in the responding patients on-treatment. C) This heatmap shows that expression of AAGAB is not correlated with known proliferation markers, suggesting its activity is independent of proliferation. D) Kaplan-Meier diagrams show AAGAB also significantly separated (log rank values) KM survival curves using the on-treatment expression values.

2.4.4 Comparison of Pre- and On-treatment Predictions of Response and Outcome

We were also keen to assess whether estimations of established prognostic signatures might be different upon treatment and if on-treatment might be more accurate. Almost all responding patients were predicted to have poor outcomes with the estimated Mammaprint,[160] PAM50,[62] or rorS[161] signatures in pre-treatment samples of the NEO cohort, whereas around half of the responding patients were predicted as good outcome using on-treatment data (Fig. 2.4.4, A, on Page 44). Overall accuracy improved by 2–8% using on- rather than pre-treatment data; however, improvement in the ability of these tests was not uniform between response classes. In this diagram red indicated the sample was classified as non-responsive, blue indicates that the sample was predicted a good outcome. Good outcome predictions for responders to neoadjuvant chemotherapy saw an aggregate increase in predictive capabilities from 11 to 44.4%, whilst poor outcome predictions for non-responders saw a moderate decrease in accuracy, 75 to 63%. These values are drawn from comparisons of the categorical classification and response status of the patients. Survival instead was considered as the end point of the study, but due to the limited number of patients and available follow up, response was considered to be a more robust and appropriate metric. Where possible, prognosis was compared to response and, in regards to the patient cohort presented, 82% of patients with good response also experienced RFS until the most recent date of available follow up. This study will not speculate how this data may extrapolate past this point but the trends suggest a positive correlation. The log rank test significantly differentiated the survival curves of the neoadjuvant responders compared to the non-responders (p-value 0.032) with a median RFS of 1045 days for the responders and 842 for the non responders. The responders only saw three patients lost to metastasis compared to ten in the non-responders.

Where clinico-pathological variables (NPI, Grade, Her2 status) were available, the hazard ratio of each feature was calculated in the pre-treatment setting at diagnosis. The hazard ratios were then compared to the calculated risks generated through the different subtyping methods and in comparison to the performance of AAGAB in the pre-treatment and on-treatment setting. None of the gene expression signatures either pre- or on-treatment or established prognostic markers from diagnostic samples (NPI, Grade, Her2 status) were significantly associated with the outcome in contrast to the remarkable performance of on-treatment measurement of AAGAB (Fig. 2.4.4, B, on Page 44) when comparing the confidence intervals of the hazard ratios. The values for this diagram are

tabulated in Table 2.3, Page 43 for all 59 paired samples from both datasets.

Pre-Treatment			
	HR	95% CI	p
NPI	2.3	-0.26-5.38	0.72
Grade	1.1	-1.98-7.40	0.81
Her2	-28.3	-53.1-27.8	0.96
Node Status	1.49	-1.18-6.11	0.74
ER Status	0.57	-1.72-3.98	0.51
Pam50	1.52	-0.22-7.14	0.63
rorS	0.09	-2.20-2.42	0.30
MammaPrint	0.06	-1.38-1.51	0.23
AAGAB	-0.1	-0.99-0.78	0.17
On-Treatment			
Pam50	0.62	-1.12-2.71	0.26
rorS	0.57	-1.82-3.0	0.29
MammaPrint	0.53	-1.07-2.11	0.24
AAGAB	-1.72	-3.02-(-0.43)	0.0021

Table 2.3: Tabulated Hazard Ratio Information This table contains the Hazard Ratio, confidence intervals and p values for each feature, for both the pre- and on-treatment tests.

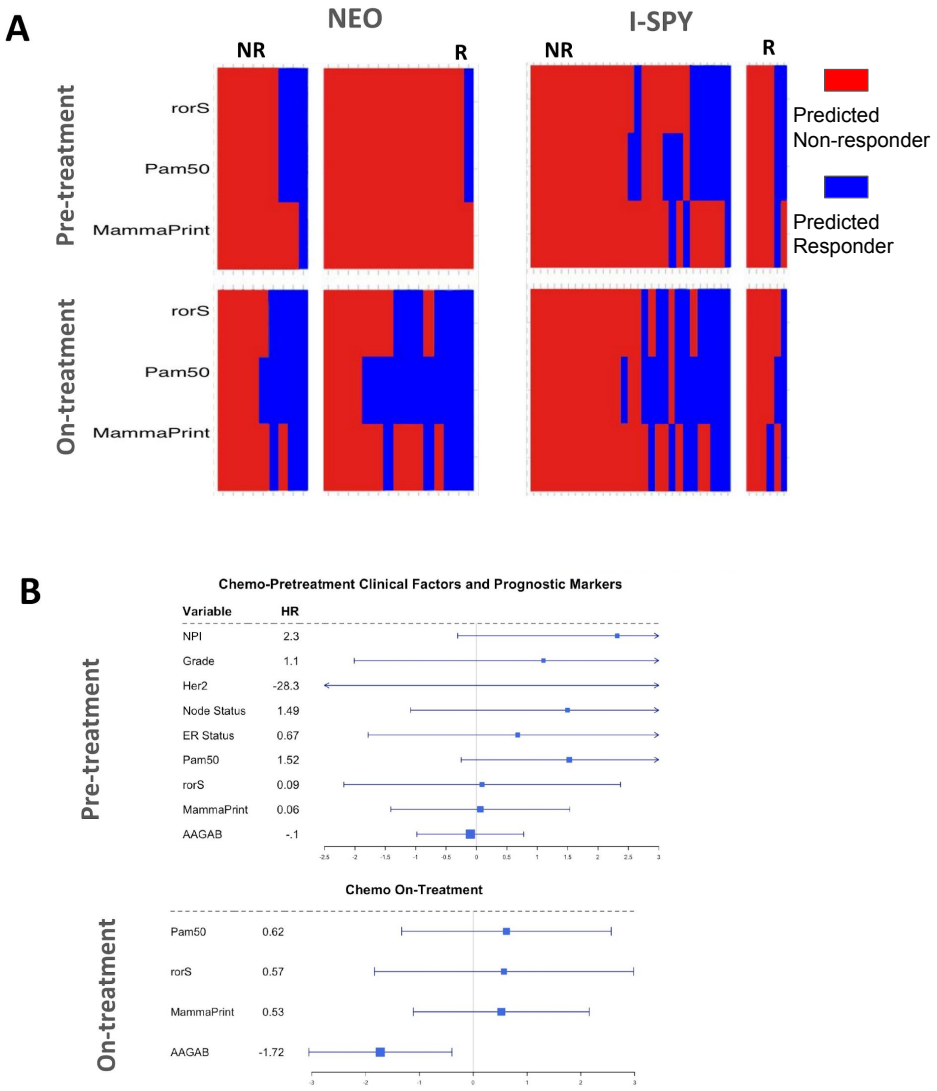


Figure 2.4.4: On-Treatment Predictive Value and Hazard Ratios. Samples are colored by predicted response, Red for non-responsive, Blue for responsive. A) On-treatment samples were more predictive of prognostic risk using existing classifiers (MammaPrint/rorS), suggesting that on-treatment expression more fully captures patient risk. This study acknowledges that Mammaprint was designed for diagnostic sample testing and seeks to compare the efficacy in a new treatment space. B) On-treatment AAGAB expression levels were the most important indicator of risk, even when comparing clinical features such as receptor or nodal status.

2.4.5 Pathway Enrichment as an Indicator of Divergent Expression

Parallel to differential gene expression between the responsive and non-responsive patient samples, GSEA analysis using common cancer adjacent hallmark pathways was performed. Figure 2.4.5, B, on Page 46 shows the results of this analysis. In concordance with the results of the LFDA and differential gene expression count, the amount of enrichment increased on-treatment. There was divergence of the pathway enrichment between the two response classes in excess of the pre-treatment samples. This strongly indicates further support for the on-treatment expression profiles being significantly different to the pre-treatment scores.

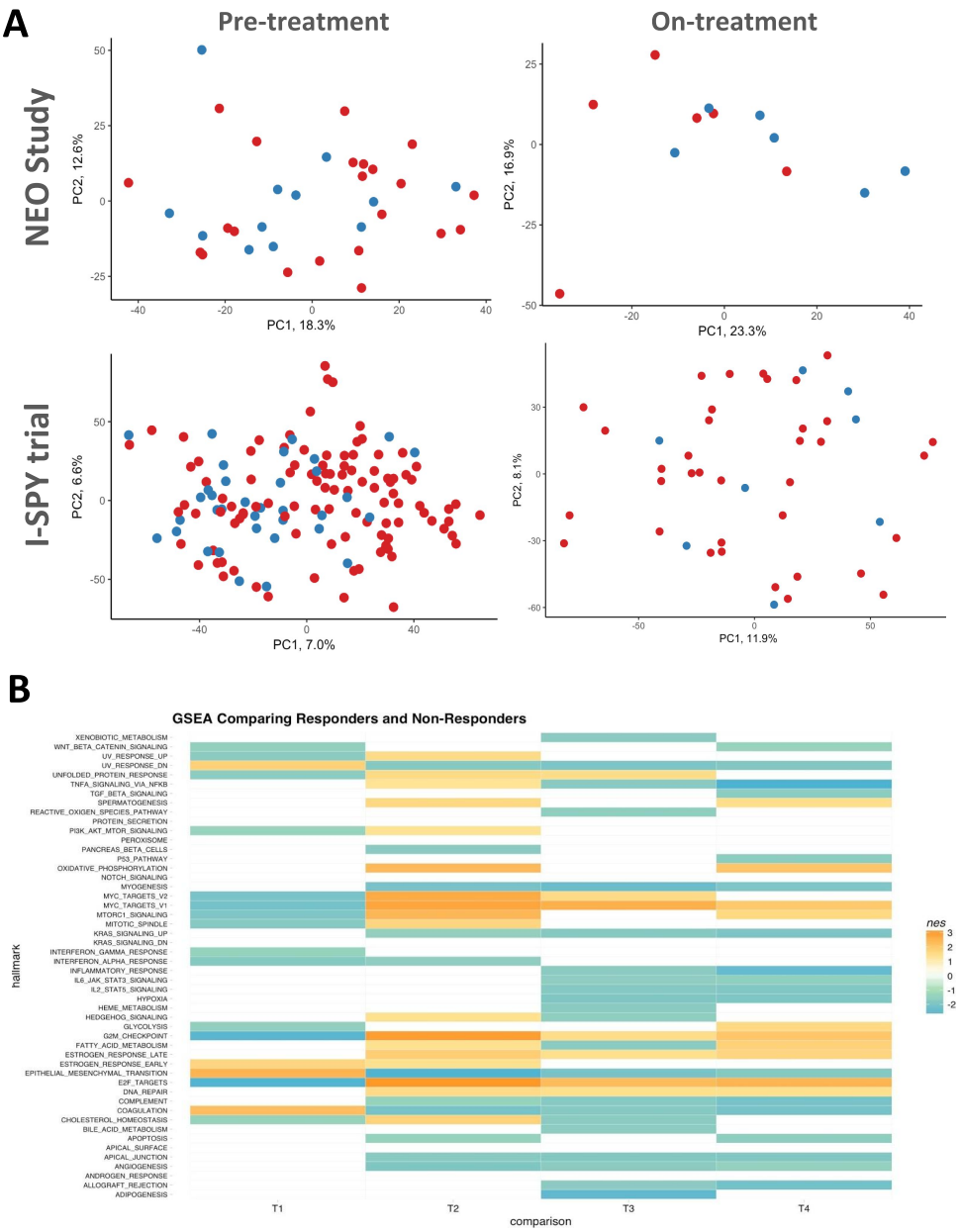


Figure 2.4.5: Further Clustering and Pathway Analysis. A) On-treatment differences have shown to be significant, but are still too subtle for fully unsupervised analysis. B) Increased enrichment of cancer hallmark pathways using GSEA shows greater activation of pathways on-treatment when compared to pre-treatment.

2.5 Discussion

Determining molecular differences between tumours to select the most effective treatment is the defining feature of precision oncology. Accurately predicting which patients will respond to treatment before exposure relies on a highly specific target. In breast cancer, ER status is a good indicator of response to endocrine treatment, but resistance, both primary and acquired, is common. Chemotherapy is an unselective treatment, relying on cancer cells growing faster than normal cells. The results presented here, along with others,[137,138] suggest on-treatment biomarkers have improved value in predicting whether tumours respond to treatment and are associated with the outcome. Changes in gene expression in sequential patient-matched samples were fairly consistent in response to chemotherapy across two independent datasets, regardless of the response status. Identifying molecular markers between responding and non-responding tumours was much more challenging. We previously demonstrated that lobular and ductal breast cancers respond to endocrine treatment in the same way, despite clear histological and molecular distinctions that are apparent and maintained on-treatment,[162] demonstrating that pre-treatment variations do not necessarily lead to differences in response. The results of this study are somewhat exploratory, rather than definitive, but further illustrate the considerable potential value of on-treatment sampling.

There are no universally agreed-upon markers predictive of response to chemotherapy, and the few that have been investigated in the neoadjuvant setting typically centre around established markers including ER, P53 HER2, and Ki-67;[163] thus, the introduction of novel biomarkers can expand the currently available clinical options for physicians. A study published over a decade ago stated that the differences in gene expression between responders and non-responders to neoadjuvant chemotherapy must be rather subtle.[142] The results presented here confirm this statement; however, our results suggest that on-treatment biomarkers may provide important information for predicting response.

As cancer is inherently a proliferative disease, measuring the change in markers of proliferation on-treatment is logical and genes like KI-67 have been demonstrated previously to be potentially a new clinical tool for disease prognosis and prediction.[97,164] It is, therefore, all the more interesting that the potentially novel biomarker identified in this study, AAGAB, is not tightly correlated with known markers of proliferation. AAGAB has primarily been studied for its role in punctate palmoplantar keratoderma[165] and the role of

adaptin in the clathrin-independent endocytosis of epidermal growth factors. The level of AAGAB was found to be prognostic of response ($p < 0.001$) in renal cancers (favourably) and in thyroid cancers (unfavourably) and expression is elevated in breast cancer, relative to the normal breast ($p < 0.001$) according to the Human Protein Atlas.[166,167] However, the exact role of AAGAB in breast cancer is currently unclear and potentially warrants further investigation. Clearly, further validation of the role of AAGAB in breast cancer is warranted and will be performed as new neoadjuvant chemotherapy datasets become available. This study supports the use and identification of genes or markers from on-treatment biopsies as a tool for improving patient response classification. We propose that the use of on-treatment samples offers valuable insight into the dynamic changes correlated with response, and submit our findings as support for continued neoadjuvant sampling, and novel biomarker generation.

2.6 Conclusion

We have identified AAGAB as a novel on-treatment biomarker for accurate prediction of pCR and outcome in patients treated with neoadjuvant chemotherapy. A semi-supervised analysis and evaluation of estimations of established molecular signatures also highlight the potential value of on-treatment biomarkers. Combining on-treatment biomarkers with known clinical prognostic factors could further improve the accuracy of response predictions and deserve further study. On-treatment expression changes in the neoadjuvant setting may offer greater possibilities for the identification and creation of more future novel biomarkers.

3 | Informing New Prognostic Decisions Through Patient Matched Tissue Gene Expression Analysis

3.1 Abstract

Background

Prediction of prognosis and response to treatment for breast cancer is heavily focused on primary tumour characteristics, even for patients with disease that has already spread to the axillary lymph nodes. Despite the presence of cancer cells in one or more lymph node being an established marker for predicting patient outcome, relatively little attention has been paid to the molecular characteristics of tumour positive lymph nodes.

Methods

Three transcriptomic datasets of primary breast cancer and patient-matched tumour content confirmed lymph nodes were generated and analysed totalling 214 samples from 98 patients to assess the relative variation between site of biopsy, different patients and the effects of treatment.

Results

Unsupervised analysis failed to differentiate tissues types, indicating strong similarity in overall expression profiles between primary tumours and patient-matched lymph node metastases. Pairwise analysis revealed 168 differentially expressed genes, with lymph node samples enriched for pathways associated with MAPK signalling, WNT deregulation tumour progression and metastasis. Lymph node samples also had significantly worse predicted outcomes with higher estimated PAM50 (p=0.0001) and Mammaprint (p=0.004) risk-of-relapse

scores. There was a 150% enrichment for basal and 172% enrichment for luminal B intrinsic subtypes in nodes compared to matched primary samples. These results are confirmed in newly available and contemporary cohorts, with 65 matched patient primary and node pairs and 18 additional technical replicate nodal pairs. On-treatment sampling showed that primary disease samples and matched nodal samples change in similar ways during neoadjuvant treatment, with uniform drops in risk of relapse scores and were distinct from the pretreatment samples after exposure to therapy.

Conclusion

This study is the largest transcriptomic analysis of patient-matched primary breast cancers and lymph node metastasis samples to-date. Lymph node samples were found to have a higher estimated risk of relapse and are assigned to worse prognosis intrinsic subtypes, but are still more tightly correlated within patients than unmatched tissues. Overall, treatment appears to affect nodes in a similar manner to primary tumours. These results illustrate the potential value of considering the molecular characteristics of lymph node samples for capturing risk beyond the primary alone.

Overview

A generalised diagram of this work, especially with regards to the generation of the most vital outcomes is presented in Fig. 3.1.1, on Page 51. This high level overview shows the flow of the paired samples from patient fold change values through to differential gene lists for response and lastly validation testing in new testing data.

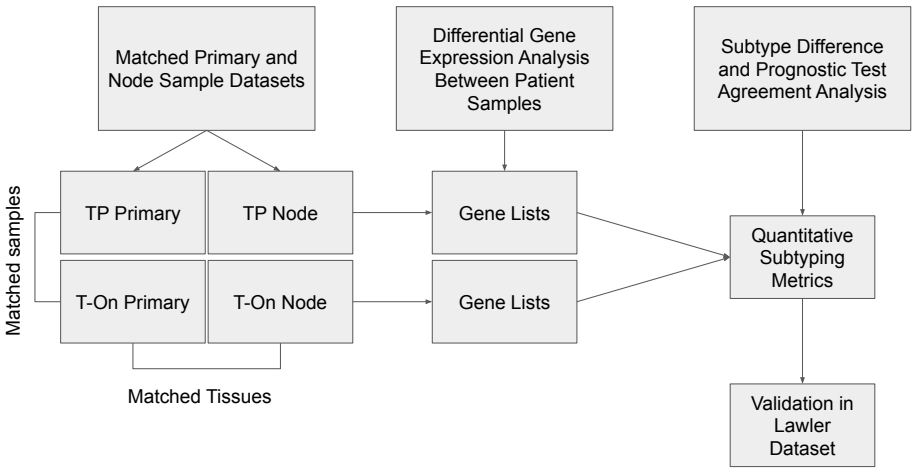


Figure 3.1.1: Overall Study Design and Workflow. This workflow diagram illustrates a high level overview of the analysis of the primary and node matched sample datasets. Starting on the left, there are the matched patient samples across time and tissue resection type, then differential gene expression analysis of the time point and tissue pairs, as well as subtype and risk prediction scores for comparative analysis. These results were then expanded and validated in a contemporary study, the Lawler trial.

3.2 Background

Lymph nodes are the most common site of breast cancer metastasis, and the spread of disease from the sentinel to subsequent nodes is an established pathway of disease progression.[168] Understanding how lymph nodes are involved in breast cancer is fundamental to predicting patient outcome, as node status is still one of the most accurate predictors of overall survival.[169] The extent of node metastasis is also crucial to proper staging of the primary tumour by defining the **T**umour size, **N**umber of involved nodes and **M**etastatic status (TNM), as the presence and number of affected nodes is part of the initial staging calculation.[170] This information is often gained through lymph node resection of some or all of the axillary lymph nodes, and by sentinel lymph node biopsy which seeks to identify the earliest involved lymph node to spare unnecessary resection to unaffected nodes, where possible.[171] Nodal involvement is important for treatment considerations and clinical decision making.[172]

A number of studies have documented the discordance between established immunohistochemistry (IHC) markers in matched primary and tumour-positive lymph node samples. ER, PR, HER2 and KI-67 have all been shown to have significant discrepancies in the expression of these biomarkers in as many as 36-54%

of patients.[173–176] Switching molecular subtypes between matched primary and brain metastasis has been observed for 45% of patients (9/20) in one study[177] and 47% (8/17) in another.[178] Discordance in classification has potentially drastic implications for clinical decision making, as treatment selection is currently made based on the known characteristics of only the primary tumour when it is almost always the relapsed tumour that is fatal. A number of gene expression studies have been performed to predict metastasis in lymph node negative patients[179–181] or compare transcriptional signatures between lymph node positive and lymph node negative tumours.[182,183]

Three studies have profiled matched primary and lymph node metastasis samples.[184–186] However, these are relatively small with patient numbers in the tens, and there is little consensus for how the differences between samples relate to established prognostic signatures used for clinical decision making or how they are affected by treatment. Recent and exhaustive meta-analysis of this space (366 studies, including 14 with genome-wide expression) have concluded that there are different expression patterns in the metastatic setting compared to primary tumour, but recognised this is still an avenue for further research, and that additional work is needed to inform new clinical interventions.[187] High resolution aCGH studies have concluded with similarly inconclusive results, indicating through cluster analysis that lymph nodes clustered closest to their matched primary, and that copy number aberrations were not notably different.[188] This suggests a high clonal relationship between primary tumour and metastatic tumour.[188] To better understand the prognostic value of these samples, further work is needed to examine consistent molecular differences between primary breast cancer and matched metastatic tissue. This study aims to expand on the body of work describing the gene expression differences between tumour biopsies by comparing the differences in relative expression and calculated prognostic risk, and using clinical follow up to ascertain the additional value to patients and clinicians.

3.3 Materials and Methods

3.3.1 Ethics Statement

Patients from the three cohorts gave informed consent to be included in the studies and were approved by the local regional ethics committees. For some samples, automatic consent was given at time of collection and other samples were consented into the trials post-collection.

3.3.2 Patients and Datasets

Dataset 1 (DS1) included 76 pretreatment samples, 24 primary samples and 52 nodal samples (of which, 22 patients had matched primary and nodes, and a further 10 patients had matched nodal samples). These samples were collected from patients of the Edinburgh Breast Unit. RNA extraction was performed via Ribo0-RNAseq, and whole transcriptome sequencing was performed with Life Sciences Ion AmpliSeq Transcriptome Human Gene Expression Kit as described previously.[189] This generated greater than 8 M reads per sample with an average of more than 90% valid reads for 14,781 targeted genes. The available annotation for each dataset is contained in Table 3.1 on Page 54. These samples were consented, collected and transported by Dr. Beatrix Elsberger, sample processing was performed by Dr. Arran Turnbull and Dr. Carlos Martinez-Perez.

Dataset 2 (DS2) included a total of 87 samples (of which, 25 had matched primary and nodes, and an additional five had matched nodes). RNA extraction was from the aqueous phase by column-based purification (RNeasy mini kit, Qiagen) and then labelled and hybridised (HumanHT-12 v4 Illumina BeadChip) according to the manufacturer's protocol (NuGEN) as previously described.[190] These samples were collected in the Edinburgh Breast Unit, the samples were processed by Dr. Laura McArthur.

Dataset 3 (DS3) represents a unique series of pre- and on-treatment matched primary and lymph node metastasis samples from HER2-positive breast cancer patients treated with trastuzumab plus chemotherapy at Ninewells Hospital in Dundee. RNA was extracted from 2x20µm FFPE tissue sections using the RNeasy FFPE kit and then analysed by Lexogen QuantSeq using single read (1x75bp) sequencing with the NextSeq 500/550 High-Output v2 (75 cycle) Kit on the NextSeq 550 platform. Data was generated from a total of 53 samples, 15 matched primary and node pretreatment samples, 13 matched primary samples across treatment, 4 matched nodal samples across treatment and 5 matched resected primary tumour and nodal samples. This dataset was originally collected and processed by Dr. Arran Turnbull and Dr. Cigdem Celli.

To differentiate between the different groups of patients according to the types of samples collected and for clarity, the following sample nomenclature will be used throughout this chapter: P1N1, relating to the patients with matched pretreatment primary and nodal samples; N1N1, meaning matched pretreatment nodal samples; P1P2, meaning to pre- and post-treatment primary tumor samples; N1N2, meaning matched pre- and post-treatment nodal samples and

P2N2, meaning matched post-treatment nodal samples.

Characteristic	Dataset1	Dataset2	Dataset3
Platform	Ampliseq	Illumina HT12v4	Lexogen QuantSeq
Gene Count	14781	17296	11425
Patients	32	49	21
Tumour Samples	24	31	33
Lymph Samples	52	56	20
Matched P1N1	22	25	15
Matched N1N1	10	5	0
Matched P1P2	X	X	13
Matched P2N2	X	X	5
Matched N1N2	X	X	4
Tumour Grade			
1	2	1	0
2	13	18	1
3	17	12	20
Oestrogen Receptor Status			
Positive	30	6	14
Negative	2	22	7
NA	X	3	0
HER2 Status			
Positive	1	17	21
Negative	1	2	0
NA	30	12	0
Number of Recurrences	16	10	7
Median Follow up (Years)	7.39	6.58	6.15

Table 3.1: Characteristics for Tumours and Patients in Datasets 1-3. Summary of all available data for Datasets 1-3, including clinical annotation and recurrence status. X indicates a feature with no information for that dataset. P1N1 relates to the patients with matched pretreatment primary and nodal samples, N1N1 matched pretreatment nodal samples, P1P2 matched pre- and post-treatment primary tumour samples, N1N2 matched pre- and post-treatment nodal samples and P2N2 matched post treatment nodal samples. HER2 status was assigned through tissue pathology staining.

3.3.3 Statistical Analysis

All statistical analysis was performed in R,[191] using packages freely available from Bioconductor[192] and CRAN, including differential expression and pathway enrichment analysis. Raw read counts were pre-processed using Limma[144] and edgeR to provide cleaned, voom-normalised expression levels for analysis. Limma and the inbuilt Bayesian methods of differential expression analysis were used to identify significantly differentially expressed genes between the tissue types, and between the treatment times where applicable.[193] Resultant

gene lists were used to identify enriched gene pathways with GO,[194] and Camera.[195] Heatmaps and multidimensional scaling plots were used to visualise the findings. The GeneFu R package[153] was used to generate estimated intrinsic subtypes and PAM50 or Mammaprint risk of relapse scores (rorS). These tests use normalised expression matrices as the input, and output a per sample assignment. Kolmogorov-Smirnov tests were used for comparing distributions, and T-tests and Wilcoxon tests were used for comparing group means, in order to quantify the significance of the between class differences. All data been submitted to NCBI GEO[196] and is awaiting acceptance to be made publicly available.

The primary method of reporting significance between two groups, or measuring the significance of the difference present between two effect populations in this study is through a combination of parametric and non-parametric tests. Standard T-tests were performed when measuring the effect between groups of unpaired samples. The five basic assumptions were always tested. First, the values are on a continuous scale (expression values). Second, the samples are (to the best of my knowledge) representative of the population as a whole, as the only constraint on the sampling process was agreeing to be part of the study and no filter was made for age or ethnicity. Third, graphically the data conforms to a normal bell-shaped distribution after processing, but is slightly long-tailed prior to normalisation. The fourth criteria is to have a reasonably large sample size is satisfied. This is satisfied as the other assumptions hold and there are sufficient samples to calculate the test statistic. Lastly, standard deviations on both sides of the mean are approximately even, suggesting homogeneity of variance. Non-parametrically, a Wilcoxon T-test was used when measuring the differences in expression of paired samples. Assumptions here were even easier to satisfy, as every sample is inherently paired and they all come from the same population, and, we are looking at within pair differences.

3.4 Results

3.4.1 Inter-Patient Gene Expression Exceeds Matched Patient Tissue Samples

The pairwise nature of the samples analysed in this study enabled assessment of the relative differences in gene expression between tumours from different patients and pairs of matched primary and lymph node metastases from the same patient, both before and on-treatment (Figure 3.4.1, A, Page 59). Comparing the

gene expression profiles of matched primary and lymph node metastasis samples revealed the expected finding that intra-patient variation is lower than inter-patient, regardless of the tissue of origin.[197] Matched primary and nodal samples taken before treatment were more highly correlated than unmatched primary or nodal samples (Figure 3.4.1, B, Page 59). This indicated that the patient, not tissue was more important for determining the overall expression profile of the sample. Additionally, this was corroborated by the data, showing that there was no significant difference between the matched primary/node and matched node/node samples. This pattern of correlation has been previously observed in untreated data, suggesting that in all cases intra-patient similarities are more significant than between.[198] Unsupervised analysis of the patient cohorts through principal component analysis or hierarchical clustering (Figure 3.4.2, Page 60) showed no major grouping of patients by tissue type. These clusters were generated using complete linkages and Euclidean distances. Identifying differentially expressed genes consistent between patient-matched pairs of primary and node metastasis samples was hampered by having three relatively small datasets to consider.

It should be noted that the numerical comparisons between datasets were drawn from scaled expression level differences, but the categorical summary differences were generated by functions of the expression level data. Improving the number of samples is important for statistical power, but the addition of new and potentially confounding heterogeneous data must be balanced with the additional significance available to the results. No direct integration of the expression level values was attempted for these cohorts, as this adds an unnecessary amount of uncertainty and removes underlying biological variance, as will be demonstrated in the next chapter of this thesis.

Limma was used to rank genes in order of differential expression, then linear models were fitted and an empirical Bayes method used to shrink the probewise sample variances and identify significantly differentially expressed genes between groups. This was performed by representing the pairwise differences between gene expression tissues as lists of contrasts to allow identification of the genes with the largest changes between categories. Following this initial step, a naive Bayes model calculates the probability of the genes belonging to either group and a decision tree model identifies the genes which pass a predetermined significance threshold (FDR > 0.05). This analysis of the relative gene expression values indicated that there were 168 genes differentially expressed between the nodes and primaries (adjusted $p<0.05$). Gene ontology and pathway enrichment (FDR=0.05) revealed 54 significantly differentiated pathways including cell cycle,

MAPK (Mitogen-activated Protein Kinase), WNT (Wingless-type MMTV integration family of genes) and IL-17 (Interleukin 17).

Despite the differential expression across the three study datasets, the expression of these genes does not appear to be consistent between the different trials. This suggests that the cut off for gene selection was too lax and that these genes were overlapping more due to their presence in each dataset, and the large biological role they play, rather than uniformity in differential pathway analysis as a result of common treatment response. These findings are reexamined later in a validation dataset.

Differentially, node metastasis samples have higher expression of a number of significant cancer-related genes including HER family members (EGFR (Epidermal Growth Factor Receptor), ERBB2 (Erb-B2 Receptor Tyrosine Kinase 2) and ERBB3 (Erb-B2 Receptor Tyrosine Kinase 3)) and MAPK members (MAPK1 (Mitogen-Activated Protein Kinase 1), MAP2K1 (Dual Specificity Mitogen-Activated Protein Kinase 1) and MAP2K2 (Dual Specificity Mitogen-Activated Protein Kinase 2)) and known proliferation markers like MYC (MYC Proto-Oncogene), IGF2 (Insulin Like Growth Factor 2) and PCNA (Proliferating Cell Nuclear Antigen). There was consistently lower expression of angiogenesis genes VEGFA (Vascular Endothelial Growth Factor A) and VEGFB (Vascular Endothelial Growth Factor B) in the nodes compared to primary samples as well as higher levels of TNF (Tumour Necrosis Factor) in nodes. Additionally, other inflammation-related genes, including IL6 (Interleukin 6), IL17B (Interleukin 17B), CXCL1 (Chemokine Ligand 1) and CCL17 (Chemokine Ligand 17), had generally lower levels of expression in the node compared to the primary. There are some studies examining the resultant gene lists of pairwise differential expression,[186] which concluded that metastatic lymph node tumours are translationally very similar to the parent tumour. Suzuki et al. note that small groups of genes are significantly expressed in the lymph node over the primary breast, and the differences are mechanistically important to metastasis; however, there was no overlap in the gene lists prepared in this study compared to that body of work. The findings presented here more closely align with Hao et al. who found differential invasion and proliferation markers in tumour positive lymph (IGFBP-5 (Insulin like Growth Factor Binding Protein 5), CD1 (Cluster of Differentiation 1) and MMP2 (Matrix Metallopeptidase 2)) through microarray data and tissue array validation.[199] Additionally Ellsworth et al. report a metastatic signature for lymph node metastasis that contains WNT2 (Wingless-type MMTV integration member 2) differential expression.[200] The aforementiond groups and the results presented in this chapter contradict other existing bodies of

work which find that prognostic markers and differential expression is held constant through metastasis.[201–204]

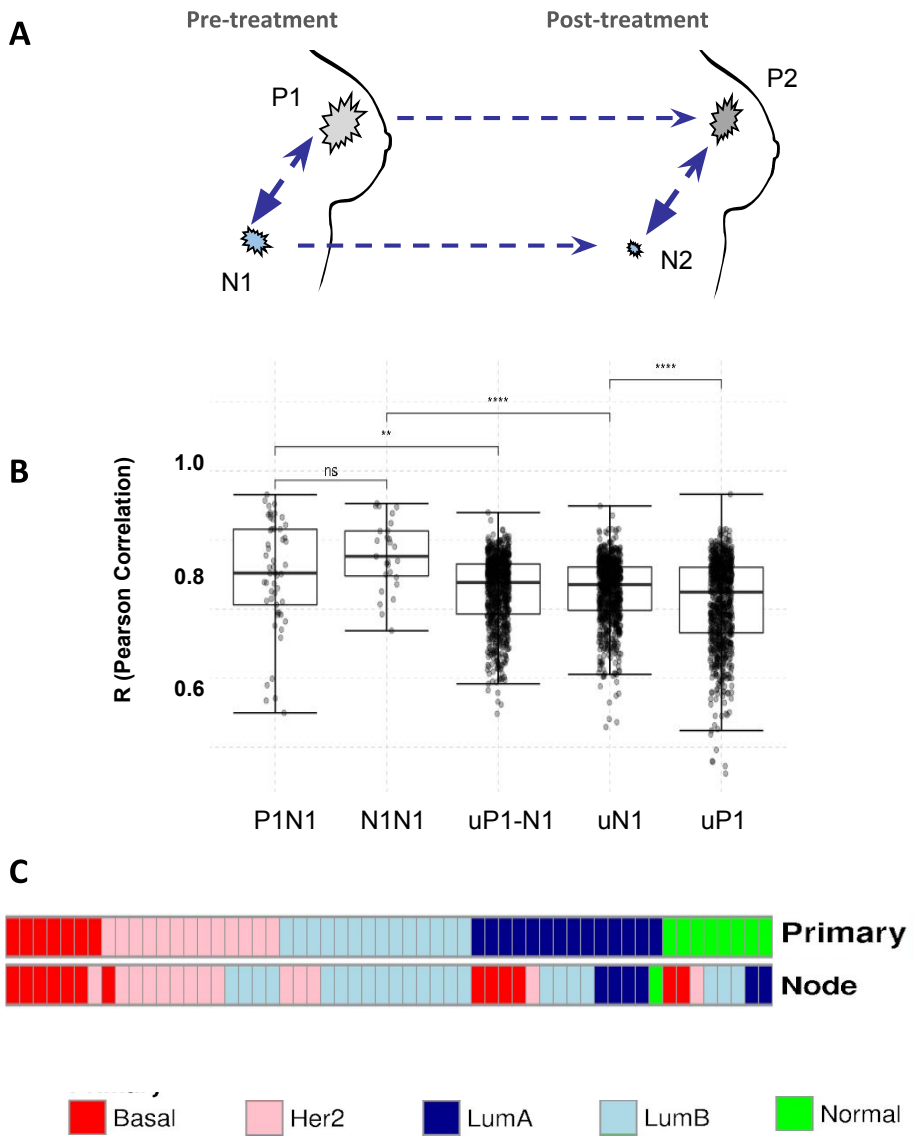


Figure 3.4.1: Study Design Diagram, Patient Correlation and Differential Expression. A) The samples in this study expand on the normal neoadjuvant design with parallel matched primary tissue and matched tumour positive lymph samples pre- and on-treatment. P1 indicates pretreatment primary tumour sample, N1 pretreatment nodal resection. P2 represents post treatment primary tumour sample, N2 post treatment nodal sample. B) The paired samples of node-node and primary-node samples pairs were significantly more correlated than samples from different patients (unpaired represented by the u prefix). c) Matched primary and node samples with their corresponding Pam50 assignments to visualise the concordance of molecular subtypes.

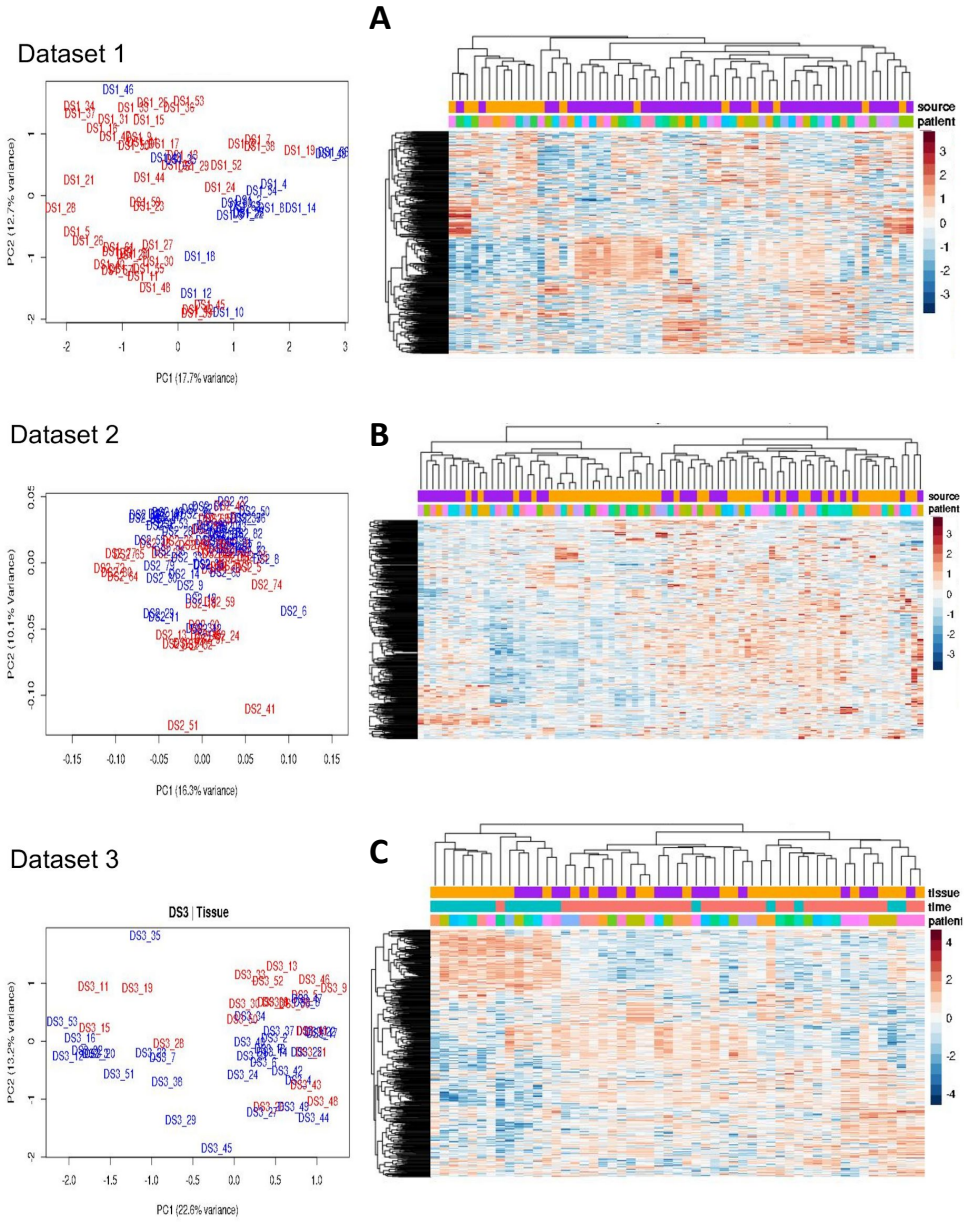


Figure 3.4.2: Unsupervised Principal Component Analysis and Heatmap Visualisation of Datasets 1-3. There is very little separation by either PCA or hierarchical clustering on unfiltered gene lists by tissue type in these three data sets A) Dataset 1, B) Dataset 2, C) Dataset 3. Suggests that the differences between tissues are subtle. Complete linkages and Euclidean distances were used to generate these clusters. Cluster robustness was not measured as this was a cursory investigation to identify the presence or absence of clear separation. Perceived groupings were checked for homogeneity of clinical features and were absent.

3.4.2 Discordance in Molecular Subtype and Prognostic Signatures

Assessment of PAM50 intrinsic subtyping[58] indicated that matched pretreatment primary and node metastasis samples (P1N1) share the same intrinsic subtype in a little over half (29/56) of patients (Figure 3.4.1, C, Page 59). However, multiple patient-matched samples of lymph node metastases prior to treatment (N1N1) were concordant in 80% (21/26) of samples, suggesting there is greater consistency in gene expression between positive lymph nodes than between matched primary and lymph node metastasis samples. Analysis of estimations of the PAM50 and MammaPrint risk of relapse scores illustrated a systematic shift between matched primary tumour and lymph node metastasis samples. The PAM50 (rorS), and MammaPrint scores were consistently higher in the node than the matched primary for 79% and 64% of patients in those tests respectively. A two-sided Kilmongorov-Smirnov test confirmed a significant difference in risk predictions between primary and nodal samples (rorS p-value = 0.0001, MammaPrint p-value = 0.004). Perhaps not surprisingly, the increases in rorS were greatest in those tumours which had relatively low scores in the primary. The rorS in the lymph node metastasis was not much higher for those patients who already had a high score in the primary tumour. Ninety five percent confidence intervals on the Loess regression support this finding (Figure 3.4.3, A, Page 62). A cumulative distribution function comparing the two distributions establish proof of a bimodal distribution (Figure 3.4.3, B, Page 62). Follow-up data for the samples is up-to-date, and there are 12 patients with a known clinical recurrence. However, for patients known to have suffered a subsequent distant metastasis, the estimated rorS scores were higher in 83% (10/12) of patients considering the lymph node score rather than the primary. Substantially increased rorS scores classified six additional patients as at high risk of recurrence based on lymph node samples but only intermediate risk based on the primary samples. Currently, there are no other studies examining the efficacy of existing diagnostic-sample-based, molecularly-derived risk profile assessment methods, like PAM50/rorS, on non-diagnostic samples (with the exception of as of yet unpublished work conducted in-institute). However, as was seen in Chapter 2 of this thesis, the techniques employed in this study appear to have improved sensitivity in the on-treatment environment, which was the impetus to test this method on locally advanced lymph tumour tissue.

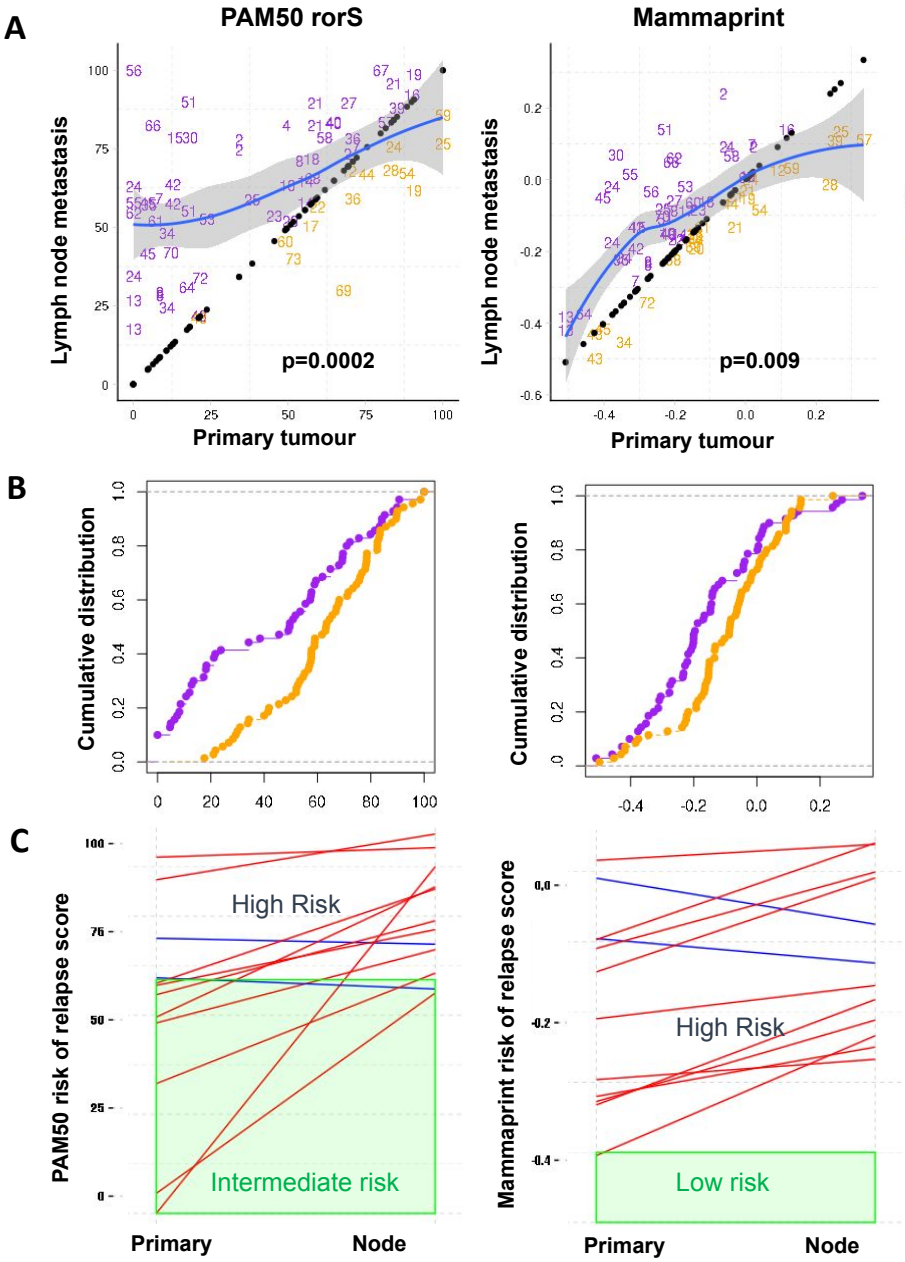


Figure 3.4.3: Statistical Comparisons of Primary BC and Matched Lymph. A) Pairwise dot plot with Loess regression line showing the relative increase in the estimated prognostic risk of the lymph sample relative to the primary tissue. B) Results of the Kilmongorov-Smirnov tests to establish significance of the distance between the distributions represented by the estimated risk of the tissue pairs. C) Change from high to lower risk in primary tumour samples relative to the matched lymph tissue is shown here to illustrate risk that is captured by the expression profile of the lymph that is absent from primary BC.

3.4.3 Validation of Pretreatment Expression Differences in Non-Local Cohort

Examination of the concurrency of the results presented in Section 3.4.2 in exogenous data was performed using data from a publicly available 2017 study of primary breast tumor with metastatic tissue.[205] To do this, the same tests used in Section 3.4.1/2 were used to compare the fraction of samples in the 2017 study with matched metastatic lymph node tissue. Sixteen matched pairs were also found in another dataset, GSE44408.[206] As can be seen from the bottom of Figure 3.4.4, Page 64 and Panel C of Figure 3.4.1, Page 59, there is a strong similarity in the representation of PAM50 subtypes in the local data and in validation, with the exception of the increased primary "Normal" samples in our own data. The forks represent the matched multiple node biopsy samples, which in the Lawler study, the 2017 study, represent technical replicate samples, hence the tight similarity between the samples. In both datasets, it is evident that when there is a change in subtype, the subsequent prognostic profile associated with the change is worse in the lymph than in the primary. It should be noted that there is a discrepancy in the overall agreement of the nodal and primary samples with 53.63% (37/69 samples) presenting differently, compared with 19.3% (17/88) differential samples in the Lawler data.

The gene lists that were extracted previously in this study were examined as pairwise differences in this cohort as well. Excluding MAP2K2 (Dual Specificity Mitogen-Activated Protein Kinase 1) and SERPINF1 (Serpin Family F Member 1) all of the genes were present and express a similar trend in relative expression, suggesting that the differential profiles of lymph may be consistently different. EGFR (Epidermal Growth Factor Receptor) and MAPK3 (Mitogen-Activated Protein Kinase 3) were consistently the most strongly differentially expressed, while some growth and proliferation markers also showed strong congruency between the studies, primarily MYC (MYC Proto-Oncogene), PCNA (Proliferating Cell Nuclear Antigen), IGF2 (Insulin like Growth Factor 2), WNT2 (Wingless-type MMTV integration member 2), CDK6 (Cell Division Protein Kinase 6) and IL6 (Interleukin 6). However, the statistical evidence is not supported by the visualisation of the underlying data as in Figure 3.4.4, Page 64. The fact that the expression levels are not more visually conforming between the studies suggests that the cut off for gene selection was too lax and that these genes were overlapping more due to their presence in each dataset, and the large biological role they play, rather than uniformity in differential pathway analysis as a result of common treatment response.

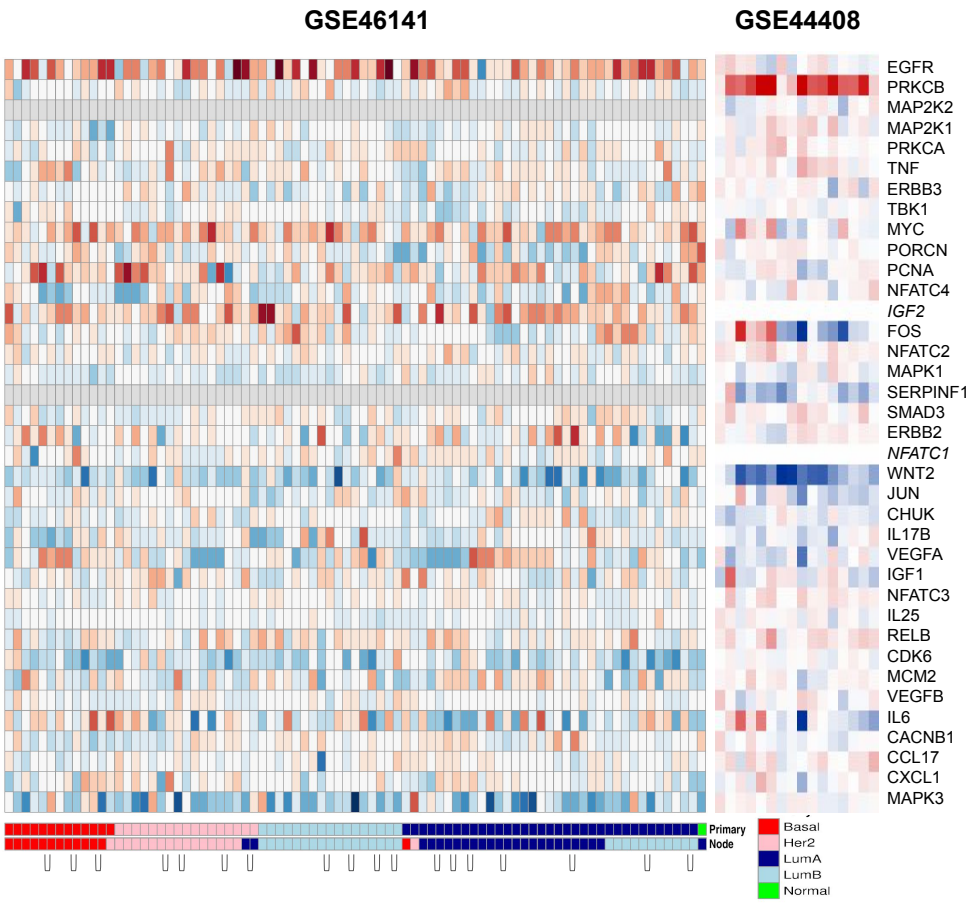


Figure 3.4.4: Pairwise Differences of Local Cohort Genes of Interest. The genes that were extracted from the iterative differential expression and pathway analysis were examined here as in the original data. Similar patterns of expression can be seen, particularly in MYC, PCNA, IGF2, WNT2, CDK6 and IL6. As in previous analysis, blue represents those genes which see lower expression in the node; red indicates higher expression in the above heatmap. The bar chart along the bottom represents the molecular subtype according to PAM50 assignment of the primary (top row) and the node (bottom row) for each sample pair, and the hair loops shows paired node samples that arise from the same patient. GSE44408 is shown on the right for comparison.

Lastly, the actual prognostic differences between the matched samples within the Lawler study show a strong parallel to our own results, see Figure 3.4.5, Page 65. The continuous rorS and MammaPrint scores indicate a systematic shift of the associated risk of the lymph when compared to the node (rorS p<0.005, MammaPrint p=0.092 (Pairwise T-Test)), and is especially true of the lower risk samples. This is in strong agreement with the results from our locally derived data, especially for the lower risk primary samples which show a greater percentage gain in predicted risk. MammaPrint reported statistically insignificant differences with p>0.05. It should also be noted the technical replicates are visibly overlapping in Figure 3.4.5 and the predicted risks are very closely matched.

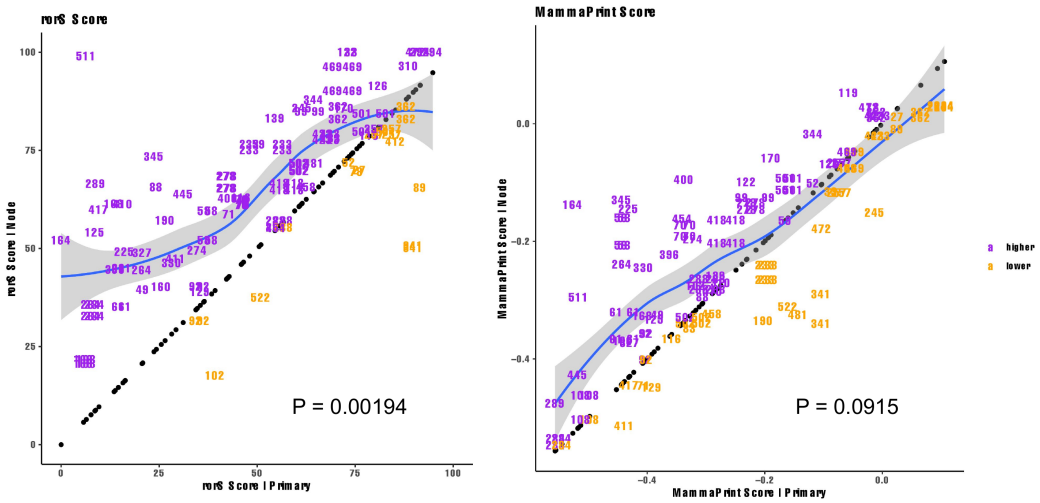


Figure 3.4.5: Pairwise Differential Risk Prediction of Primary vs. Nodal Tissue. The results of the pairwise comparison, in the Lawler dataset, between the predicted risk in the node and the associated primary tissue follows much the same pattern as in our local dataset, with significant differences correlating to the tissue of origin. Purple represents the samples with a higher rorS or MammaPrint score in the node; the orange samples conversely represent the samples with a higher risk presentation in the primary tumour.

3.4.4 On-treatment Gene Expression Differences Between Primary Tumours and Paired Lymph Node Metastases are Primarily Maintained

Unsupervised principal component analysis of Dataset 3 clearly delineated samples by time, suggesting that gene expression is more affected by treatment than the site from which the tissue originated (Figure 3.4.6, A, Page 67). Cluster analy-

sis of the five hundred most variable genes supports this finding, with the highest branching of the dendrogram enriched for on-treatment samples, with a significantly different pattern of expression to the other samples. Additionally, there were no notable groups of patient or tissue clusters visible on a heatmap of total linkage and euclidean distance (Figure 3.4.6, B, Page 67). In addition, the correlation of samples is markedly changed on-treatment. The matched patient samples of pre-treatment to on-treatment primaries and nodes showed lowered average correlation compared to the pretreatment pairings suggesting a strong profile expression change with treatment (Figure 3.4.6, C, Page 67). Clustering of the pairwise change values from pretreatment to on-treatment for the primary pairs and nodal pairs revealed that only seven of the differentially enriched genes were retained on-treatment and there was no clustering of patient by tissue (Figure 3.4.6, D, Page 67).

Assigning molecular subtypes to the samples from the small number of patients with primary tumour and lymph node metastasis before and after treatment revealed that subtype was largely concordant (75%, 6/8) between primary and lymph node at the same time point (consistent with the clustering and differential expression analysis), but was often discordant following treatment (63%, 5/8, Figure 3.4.6, E, Page 67). As can also be seen from Figure 3.4.6 all four samples from one patient were classified as HER2-enriched, whilst the other three patients switched from HER2-enriched or luminal B to luminal A or normal-like during treatment. Two of the patients would have been classified as HER2-enriched based upon the lymph node sample, but normal-like and luminal B based upon the primary tumour prior to treatment. The estimated risk of relapse scores were significantly (Wilcoxon p-value, 0.0189) lower post-treatment than pretreatment. Three of the four patients with matched primary and lymph node samples before treatment had reductions post-treatment in the scores of both tissues (Figure 3.4.6, F, Page 67). This result is not surprising, as down-regulation of genes involved with proliferation is expected during treatment and is a prominent feature of prognostic signatures.

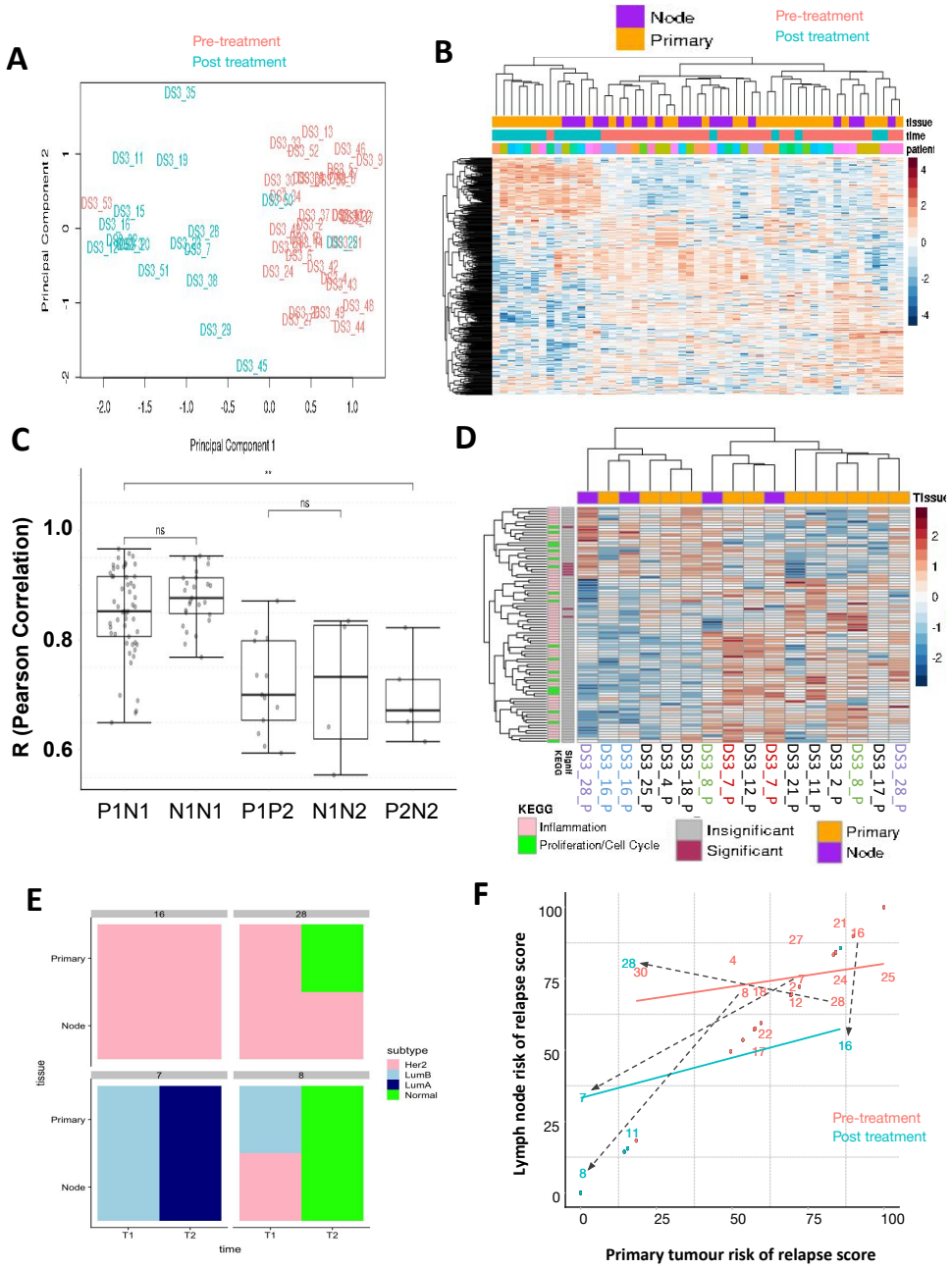


Figure 3.4.6: Expression Profile Progression with Treatment in Matched Primary/Lymph Samples. A) PCA diagram illustrating clear separation of the pre- from on-treatment samples. B) Heatmap with total linkage and euclidean distance of the 500 most variably expressed genes on- to pretreatment with clear isolation on the basis of biopsy time. C) Correlation box plot of matched patient samples drops on-treatment, suggesting diverging expression profiles with therapy. P1 indicates pretreatment primary, P2 post treatment, N1 pretreatment node, N2 post treatment. D) The heatmap as in A shows that there are far fewer differentially expressed genes between nodes/primary tissue pairs on-treatment. E) This tile plot shows that subtype concordance is more similar across tissue than treatment time in matched samples. F) This line plot shows the reduction in estimated prognostic risk is uniform between the tissue types on treatment.

3.5 Discussion

The presence of cancer in lymph nodes has long been established as an adverse prognostic indicator, and previous studies have observed discordance in individual receptor levels between primary breast tumours and lymph node metastasis. However, to date, this is the largest study to take an unbiased data-driven gene expression comparison between primary breast tumours and lymph node metastasis. The results presented here demonstrate that gene expression profiles of patient-matched primary breast cancers and lymph node metastasis samples are much more similar than tumours and lymph node samples from unrelated individuals. Previous studies have made similar statements on the expression profiles of matched patient samples,[202–204] but did not examine changes on-treatment. However, there are consistent gene expression differences between primary tumours and lymph node metastases. The difference in general expression level profiles between the tissue types is a logical result given that lymph tumour tissue should display increased pro-metastatic and infiltrative characteristics, an effect observed in other studies[186] with varying degrees of overlap to this work.[199,200] This change in expression profile, and the tumour's ability to adapt to and leave the local site of the primary tumour, go hand-in-hand.

These results demonstrate that transcriptomic analysis of pairwise primary tumours and lymph node metastasis samples can contain important new information about disease progression. This leads to a potential improvement in the ability to accurately capture the prognostic risk presented by primary breast cancer and associated lymph metastases on a per patient level through new or existing methods like PAM50.[207] Changes in the calculated risk or molecular subtype may have implications for treatment response. Additionally, this risk can equally be captured by any tumour positive lymph tissue sample, as the differences between matched primary and node were larger than between matched nodes of the same patient. This indicated that, while the lymph may present at a higher risk than the primary, the expression profile and the risk of associated nodes is more uniform.

The response to therapy was largely consistent in the two tissue types. On-treatment, gene expression differences due to treatment were much more prominent than differences between tissue types, which were maintained. It is encouraging that lymph node metastasis samples respond to treatment in a similar way to the matched primary. There are parallels with our previous study, which demonstrated highly similar transcriptional response to endocrine

treatment in ductal and lobular carcinomas despite clear histological and molecular differences between the subtype pre- and post-treatment.[162] While there is an apparent scarcity of these samples, we hope that the results demonstrate the value of further work into these matched tissue samples, as this could facilitate a more in-depth and powered study into the differences between tissue types of patients who do and do not respond to treatment.

The same relative increase of differential risk associated with the lymph node samples were seen in the non-local data validating the observations made on our own local cohorts. The lymph tissue biopsies show strong similarity for risk stratification in terms of comparative subtype-associated risk, differential gene expression and the significant difference in the continuous prognostic risk compared to the primary breast tissue. This cohort almost doubles the number of available samples, and is strong supporting evidence, barring the missing follow up data for the study, that the patterns of change which were observed are statistically significant and not a random occurrence.

3.6 Conclusions

Through the combination of correlation and differential expression analysis, along with classification and prognostic signature evaluation, it has been determined that matched lymph node metastasis samples have more aggressive gene expression profiles and higher predicted risk of relapse than patient-matched primary breast cancers. This increased predicted risk can potentially be translated to an improved ability to predict recurrences in patients over the primary tumour biopsy results. Tumour expression profiles were markedly different post-treatment, and changes due to therapy were much greater than differences between tissue site. This study suggests that, where applicable, matched node biopsies may be of additional value to clinicians and patients as an additional tool for risk stratification and treatment decisions.

4 | Evaluation of Approaches to Integrate Sequential Pre- and On-treatment Patient-Matched Breast Cancer Datasets

4.1 Abstract

Background

Neoadjuvant therapy represents a unique opportunity to monitor tumour expression level changes and response to treatment in a similar time frame to tumour evolution. Changes in the transcriptomic landscape of cancer on-treatment have already been shown to correlate with response and outcome in labelled patient data from modestly sized studies. Can more significant findings be reached by combining several contemporary datasets, and will this allow us to further the goal of personalised medicine?

Methods

Five transcriptomic datasets of primary breast cancer were collected and combined using *ComBat* to create a unified expression dataset with matched annotation data. Results of each *ComBat* integration method were compared with unintegrated and separately analysed data to establish discordance, and were also compared to parallel methods to evaluate performance.

Results

Integration of patient-matched sequential sample datasets proved remarkably complex. Attempts with existing techniques exposed new challenges to large scale integration. Biological and systematic covariates were evaluated to improve batch correction methodologies; however, integration of these datasets failed to return concordant values for expression, intrinsic subtype or calculated risk scores (agreement scores for calculated subtype ranged from 19-30%). Additionally, matched patient correlations were low (median 0.62), and differential expression between the altered and unaltered data was non-overlapping (mean 30.2% shared pathways), suggesting systematic changes to the expression profiles post-integration.

Conclusion

This study evaluated methodologies for the integration of multiple sequentially sampled datasets to create the largest uniformly annotated transcriptional dataset of sequentially sampled neo-adjuvant primary breast cancer. However, unifying transcriptional data across multiple studies proved an insurmountable task to accomplish, while still maintaining the underlying biological variation. Alternatively, considering only the pairwise differences within each dataset allowed for cross-study comparisons and improved analysis. This resulted in an expression list object of similarly normalised data, with unified annotation data for analysis and is presented to provide opportunities for on-treatment biomarker identification and validation. This solution provides improved statistical power for the identification of pan-treatment trends in breast cancer and a pragmatic and workable solution for comparing and combining datasets.

Overview

A generalised diagram of this work, especially with regards to the generation of the most vital outcomes is presented in Fig. 4.1.1, on Page 73. This highlights the flow of data from the individual datasets, through multiple different integration methodologies while measuring the changes to the original data, and finally through different subtype and risk assessment test agreement levels.

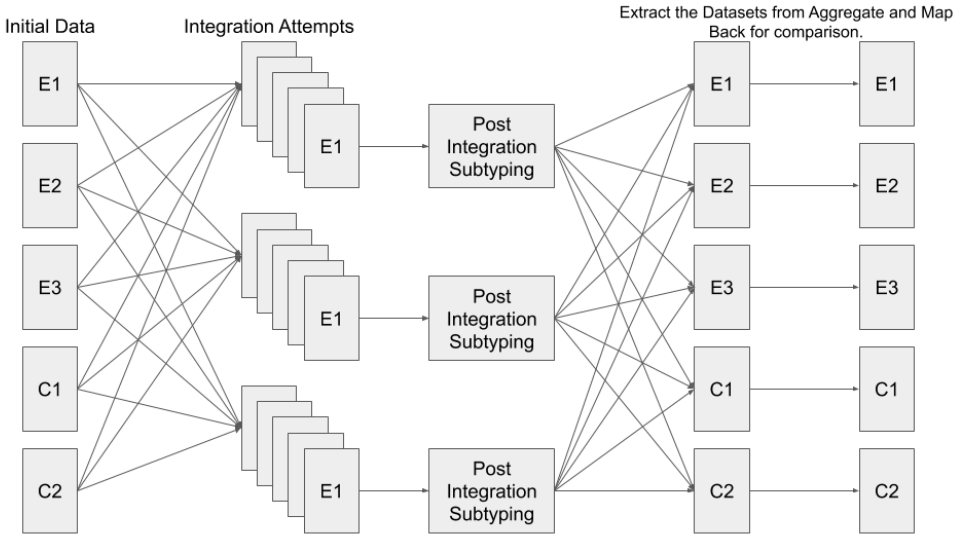


Figure 4.1.1: Overall Study Design and Workflow. This diagram shows the paths the data has been put through to compare the efficacy of each integration method. The individual datasets are combined using different integration methods, then the resultant integrated datasets are subtyped and scored for prognostic risk. The starting datasets are retrieved from the composite dataframes and compared to the starting data to evaluate the performance of each integration method.

4.2 Background

There is an abundance of pretreatment only breast cancer gene expression data sets like with 50-300 patients like the NKI,[159] with a few larger exceptions like METABRIC[208] (n=2,506) the breast cancer specific section of the TCGA[209] (n=987). Recently, tools and packages have been made available for accessing this type of data from curated repositories (MetaGXData, specifically Breast,[73] and curateTCGA[210]). Additionally, there are tools leveraging the vast amounts of well annotated data on NCBI GEO to allow for the bespoke creation and testing of meta-datasets like ImaGEO,[211] which facilitate the acquisition and probing of specific, relevant GEO accessions for analysis. While it is broadly understood that more samples and a more representative cohort improves the significance of results, these are examples of meta-analysis, but do not demonstrate true data integration. These meta-datasets increase sample size and representativeness of the data, but, fundamentally, only test a hypothesis or assertion in several datasets independently, pool the results and report the aggregate. Integration of -omic data would assist in a new and improved understanding of the underlying biological processes and mechanisms of disease, like cancer,[212] but is a non-trivial undertaking. This has also never truly been undertaken for sequential or neoadjuvant studies, even though large datasets of well annotated

pretreatment only data have resulted in several novel approaches for breast cancer subtyping and risk assesment (70 gene signature, IC10, respectively), including FDA approval and expansive clinical trials.[213] This 70 gene signature is more commonly known as MammaPrint, and is a prognostic test designed for use in diagnostic samples to help make treatment decisions for BC patients. IC10 is a molecular subtyping algorithm which classifies breast cancer based on the activation or suppression of upstream cancer drivers and denotes classes with distinct prognostic risks. As of yet there are no large scale datasets for matched patient samples or window studies.

Neoadjuvant therapy for primary breast cancer is showing significant results for patient care and outcome, including higher rates of breast conserving surgery without an increase in long term distant recurrence.[214] We have already shown that there are distinct transcriptional differences on treatment that can be used to differentiate responsive and non-responsive primary breast cancer in neoadjuvant chemotherapy[215] in two modestly-sized cohorts. Additionally, this work strongly suggests these samples being of increased value to existing prognostic tests, however, this required a larger sample size to validate. Additionally, Turnbull et al. previously showed that, in a similar fashion, samples from neoadjuvantly treated endocrine therapy patients could be used to create novel biomarkers for the prediction of response to aromatase inhibitors.[216] These findings suggest that the fold changes seen on-treatment in the neoadjuvant setting for sequentially sampled breast cancer may be of inherent value for the treatment and management of breast cancer. Turnbull et al. also performed the first and only successful integration of multiple on-treatment datasets.[217] That 2012 study examined the integration of two different platforms with very similar cohorts and had the significant advantage of replicate patient samples across the platforms to validate that integration did not distort the expression values.[217]

Neoadjuvant trials are, however, comparatively scarce, and there are additional factors to consider for these samples, including the increased cost for multiple biopsies and the associated increased patient stress. There is potential for justifying these risks with significantly improved insights into risk stratification and prognosis, but integration is required for a sufficiently large cohort. Integrative analysis makes the basic assumption that effect size is constant between neoadjuvant matched patient trials and standard diagnostic examinations. It may be the case that neoadjuvant longitudinal studies have greater signal-to-noise ratio, this effect is a presupposition of this study. Here, we propose the trial of integration techniques for patient-matched sequential samples of primary

breast cancer to ascertain the viability of cross-study combination. These methods will attempt to retain the composition and biological variation of the pre- and post-treatment samples, while also leaving the matched pre- to on-treatment gene expression level changes intact. Figure 4.2.1, Page 75, shows the concept of an "independent" analysis of multiple datasets and the integrative approach.

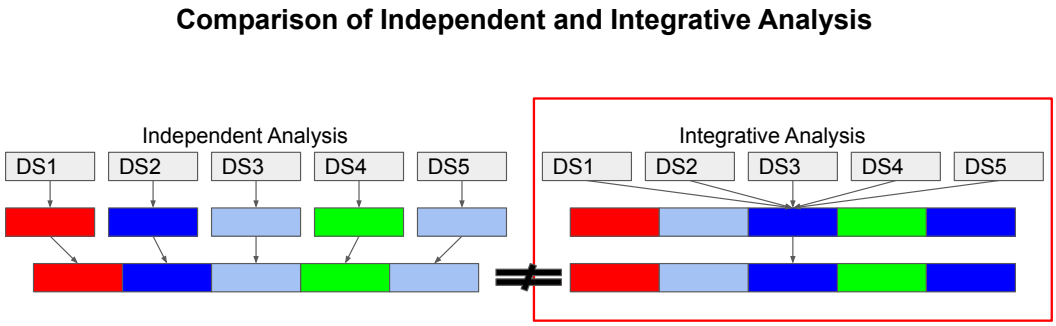


Figure 4.2.1: Schematic Diagram Illustrating Independent vs. Integrative Analysis. Independent, or meta-analysis, seeks to test one dataset and make inferences with data from another, or, in this case, by testing multiple datasets simultaneously and comparing the results. Integrative analysis attempts to first unify the data, before probing it. This method of aggregation can lead to differences in the presentation of the data. This diagram is intended to illustrate the different approaches to integration of data, either through the meta-analysis of individual examinations (left) or by the aggregate interrogation of combined studies (right).

Figure 4.2.2, on Page 76 shows the mean-centred gene expression distributions for every dataset (A); the sub distributions of treatment-by-time point as box plots (B); and the kernel densities (C) of the five datasets to be explored in this chapter. This diagram should help to show that, even in a relatively small number of datasets, there can be large differences in the expression levels of genes between datasets. It should be noted that even within a single data set there can be clear subdistributions of the data, which may stem from technical or clinical factors. This diagram highlights the complexity and depth of the analysis required to integrate these different datasets efficiently. To disturb the underlying distribution would alter relative gene expression values, which in turn would make subsequent analysis invalid, as changes in gene expression would no longer be linked to biological differences but to integration methodologies instead.

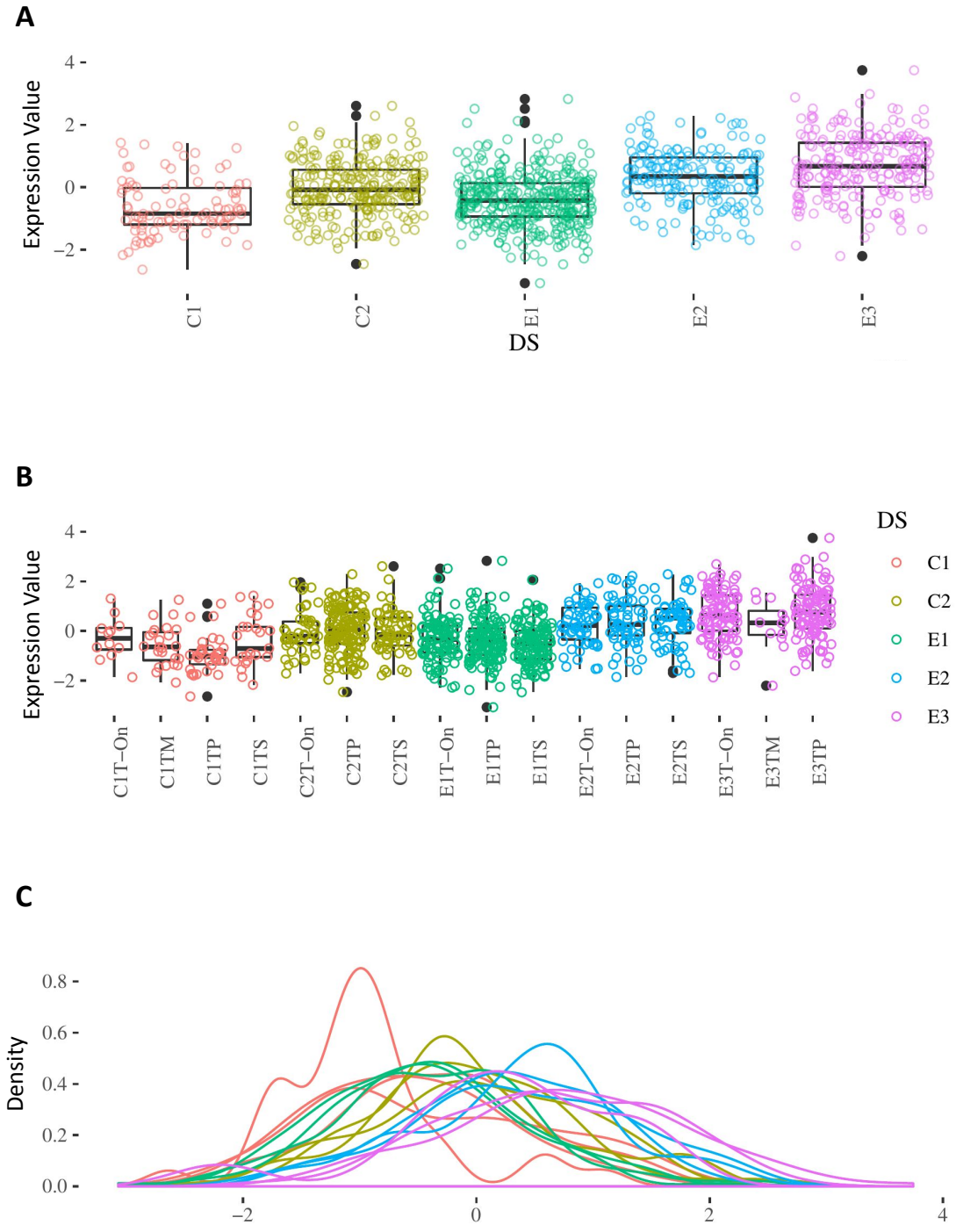


Figure 4.2.2: Pre-Integration Expression Distributions and Subdistributions. A) There are clear differences in the overall expression profiles of the five different cohorts. B) When accounting for the study of origin as well as the time points, significant splits in the expression profiles appear at each on-treatment interval. This suggests that treatment is having a profound effect on the overall expression of the samples, and that each time point represents a new unique distribution. C) This effect can be observed in the kernel densities and highlights the non-uniformity of the data and the number of distinct distributions that must be shifted by ComBat.

4.2.1 Integration Methods for Transcriptomic Data

Methods for the integration of continuous microarray expression data include Surrogate Variable Analysis,[218] Bayesian Factor Regression Models,[219] Factor Analysis[220] and Array Generation-based gene Centering.[221] These methods all seek to shift the distributions of expression towards one another to achieve integration. *ComBat*[222] has been established as the most widely used method for batch correction and integration of microarray data,[223–225] and will be utilised in this study as the primary tool for integration. Recently, Cross Platform Normalisation (XPN) has been shown to be a reliable method of microarray integration for the purposes of machine learning on microarray and RNAseq data,[226] but this is a new result and XPN is designed for only two "batches" at a time. *ComBat* uses an empirical Bayesian method with a combination of additive and multiplicative terms to calculate and remove the batch effect from the gene expression data. Due to the over-representation of *ComBat* in the literature of microarray and transcriptomic data integration, it is the logical starting place for attempting to correct the effects of batch introduced by the high number of technical and biological confounding factors in this data. This technology does, however, rely on the assumption that the expression data is in some way effected by these factors,[222] this constraint will be examined for its effect on the resultant integration. We will discuss the steps required to integrate the disparate datasets and the decisions made to analyse the effects of the most important covariates, including dataset, biopsy time, treatment type and calculated intrinsic subtype.

4.3 Materials and Methods

This study is comprised of five datasets of either neoadjuvantly chemotherapy treated patients with matched pretreatment, on-treatment and surgical biopsy samples, or aromatase inhibitor treated patients with similarly matching samples. The datasets are all publicly available; details of the total study size, treatment architecture, biopsy schedule and platform are contained in Table 4.1 Page 79, as well as accession numbers for reference studies. These studies include: Miller et al., 2012 (data accessible at NCBI GEO database, accession GSE20181), Arthur et al., 2014 (data accessible at NCBI GEO database, accession GSE55374), Turnbull et al., 2015 (data accessible at NCBI GEO database, accession GSE59515), Magbanua et al., 2015 (data accessible at NCBI GEO database, accession GSE32603), Bownes et al., 2019 (data accessible at NCBI GEO database, accession GSE122630) and Ellis et al., 2017 (data accessible at NCBI GEO database, accession GSE87411). There is an imbalance of the

study composition between the five cohorts, which is reflected in the number of matched paired samples and the number of possible matches, Table 4.1, Page 79. Cross study comparisons of time points are only made where direct comparisons exist in both cohorts. Additionally, the numbers of matched pairs for each study are tabulated in Table 4.1, Page 79. This repository includes the full annotation for the combined cohorts (representing 28 features in its entirety), including sample and patient ID, ER, PR and HER2 status, clinical response, nodal status, tumour grade, T,N,M classifications, RFS, treatment type and schedule, processing platform and dataset of origin, a new study ID variable, molecular classifications (such as scmod1/2, PAM50, ssp2003/6, IC10), rorS risk and score and MammaPrint risk and score. These features describe the patient in their clinical and prognostic presentation, as well as many derived features, like the molecular subtyping values that allow for granular analysis of the data. These studies were chosen because they are amongst the largest and most well annotated publicly available patient-matched sequential biopsy studies. Special notice should be paid to Dataset E2, which contains duplicated samples from Dataset E1. However, they have been separately processed, and on different platforms (Affymetrix), which means the overall patient number is reduced. These samples will still be analysed to identify the changes that occur as part of integration. This knowledge was not known at the start of the analysis, however, after considerable discussion on the overall impact of these samples on the analysis of this data, the decision was made to continue with the work uninterrupted. Furthermore, figures 4.2.2, Page 76 and 4.4.2, Page 93 support this decision by showing that the overlapping samples do not appear to group together or have similar expression profiles. This is primarily due to the fact that the changes that are most important to this analysis are all calculated on a per sample basis and, as these samples appear superficially different, the changes can be considered individually.

Dataset	TP	T-On	TM	TS	Total	GEO Accession ID	Study Structure	Treatment	Platform
E1	131	108	0	132	371	GSE20181, GSE55374, GSE59516	Pretreatment, two weeks, mid chemo, surgical biopsy	Anastrozole, Letrozole	Illumina HT-12 V4, Affymetrix U133A
E2	58	58	0	60	176	GSE32603	Pretreatment, 24-72 hours, surgical biopsy	Letrozole	Affymetrix U133A
E3	109	77	12	0	198	GSE59515	Pretreatment two weeks, three months	Anastrozole, exemestane, letrozole, FEC, Docetaxel	Agilent Microarray 4X44K
C1	34	12	24	25	95	GSE122630	Pretreatment, 10-14 days, surgical biopsy	FEC, Docetaxel	ION Ampliseq
C2	141	45	0	62	248	GSE87411	Pretreatment, 2-4 weeks	Anthracycline, taxane	HAQQLAB_HUMAN_40986
Total	473	300	36	279	1088				

Table 4.1: Available Samples and Study Features for Each Dataset. Not all samples were available at every time point. This table shows the available samples at each biopsy time, and an X signifies no biopsies at that time point. TP represents the Pretreatment Sample, T-On is two-weeks on therapy, TM is mid-chemo and TS is the surgical biopsy sample.

4.3.1 Data Selection and Acquisition

Collection of the data was primarily performed with GEOquery[227] from inside R. GEOquery collected the expression data, annotation, and GPL (GEO Platform) information for all datasets except Dataset E1 which was not hosted on NCBI GEO. E1 is hosted locally and is comprised of three datasets, two partially overlapping studies of Illumina and Affymetrix data, a description of this dataset is detailed in Turnbull et al.[228] Preprocessing and normalisation of the data was handled in R,[145] with the standard library and Bioconductor[229] packages Limma[230] and edgeR.[231,232] Subtyping and risk assessment were performed by the geneFu package.[233] Unsupervised cluster analysis was performed with base R `prcomp`, and all visualisations were handled by ggplot2.[234] Data manipulation and cleaning was performed primarily within tidyverse,[235] data packaging and function wrapping was done using base R[145] and Biobase,[229] as well as XDE[236] for custom S4 data types, and oxygen2[237] for documentation.

4.3.2 Preprocessing, Normalisation, and Analysis of Expression Data

Each dataset was initially cleaned of low coverage and missing probes and preprocessed individually. Read counts for each probe were converted to counts per million and filtered for genes with expression in at least half of the samples. This was done to avoid inclusion of genes with no recorded expression level data (NAs), and was performed entirely in R by Limma[230] and edgeR.[231,232] Library counts were summed for each sample to ensure that each passed a per-dataset threshold and was not significantly larger or smaller than its contemporaries. The datasets were all voom normalised on the same features, Law et al. describe this process in great detail and establish voom as a viable method for heterogeneous data processing.[238] E1, E2, E3, C1 and C2 started with 11212 15311, 21089, 13099 and 14112 genes, respectively, and ended with 7072, 12513, 15513, 12633 and 13411 genes, respectively. Initial subtypes and risk scores for IC10,[154–156] PAM50,[156] scmod1[157] and scmod2,[158] ssp2003[154] and ssp2006,[155] MammaPrint[159] and rorS[156] continuous and categorical risk values were calculated with geneFu. MammaPrint, rorS, PAM50 and IC10 have been described previously, however ssp2003/6 and scmod1/2 have not. The ssp2003/6 tests are methods of calculating BC subtypes first proposed in 2003 then updated in 2006, and have good accordance to clinical assignments made by ER IHC, HER2 IHC/FISH. Scmod1/2 are also gene centroid mapping utilities based off lists of 726 and 663 genes correlated to ESR1, HER2 and AURKA that predict HER2+ ER+/-/HER2-/High/Low subtypes. These tests were all designed for samples in the diagnostic/prognostic setting

and are thus exploratory for measuring these values in the post treatment setting. However, preliminary results such as those in chapter two of this thesis show that there may be added functionality when applied to calculating risk assessment by leveraging the on-treatment changes.

Changes in the subtype composition results of these geneFu tests were made using the `compareDF`[239] package in R and tabulated. GeneFu is a centroid mapping distance based subtyping algorithm. The program will read in flat tabulated data annotated with samples per column and genes, usually Entrez Gene Id's, as the rows. It will transform the data into a n-dimensional coordinate and then find the distance between the point plotted for each sample and the subtype centroids. The subtype with the least distance to the sample is chosen as the assignment for that sample. These changes represent categorical state changes in the presentation of the samples as a result of integration. Limma was also used for differential expression analysis, and lists of differentially expressed genes were compared pre and post integration for each dataset to identify changes in the driving genes per cohort. Models were trained to try and predict datasets of origin from the integrated data as an additional method of identifying data structure perturbance. Accuracy scores may be reported as balanced F1 scores, this has been calculated as:

$$F_1 = \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

This is a balanced metric between Precision and Recall and is a sensitive metric to imbalanced class sizes as are present in parts of this analysis.

4.3.3 The Origins of Technical and Biological Variance

There are numerous confounding factors to consider for cross-platform comparisons, each of which must be accounted for and minimised during the integration process, see Figure 4.3.1, Page 82. These factors, including platform and treatment type, may occlude common biological similarities between datasets and must be corrected for. The correction process itself may also incur additional noise from the new transformations, another consideration for this process. The impact of these factors will be analysed post-integration in three key ways; with single time point pretreatment only sample integration, non-informed uncorrected joining methods, and *ComBat* correction with additional covariate information for each of these artefacts.

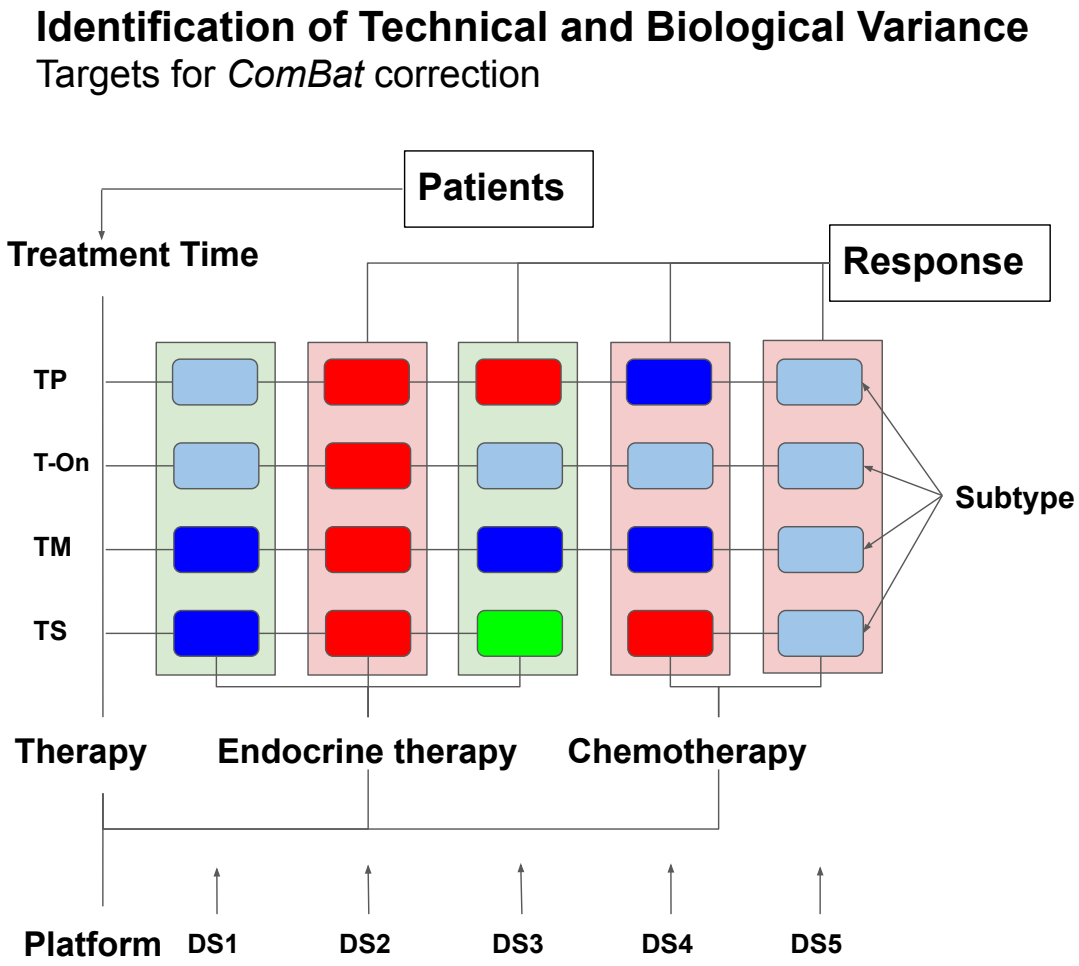


Figure 4.3.1: Concept Map of the Various Biological and Technical Factors to Consider for Integration. This diagram highlights some of the considered factors for covariates of ComBat analysis due to their intrinsic ability to convolute cross study comparisons. In this example, Platform, Therapy and Treatment Time are considered batch effects and response and subtype are biological covariates. Not highlighted here is the importance of cell-type heterogeneity, which is not further explored.

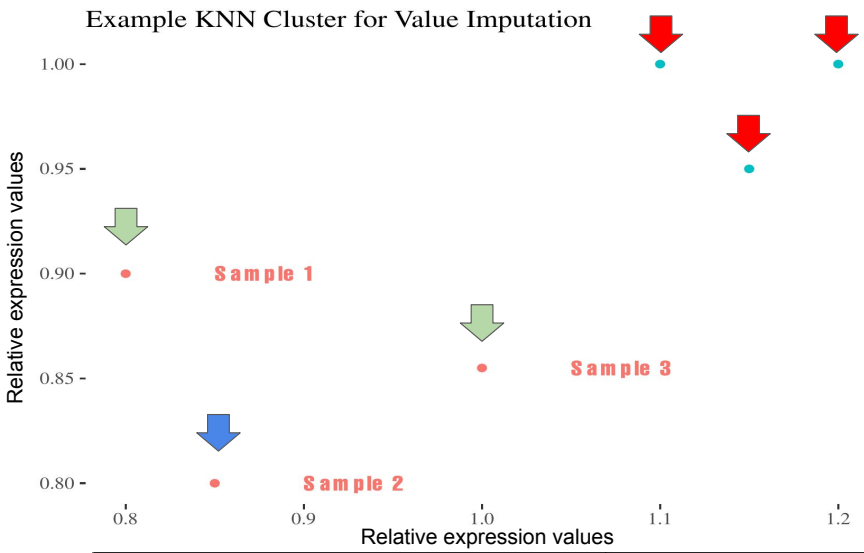
4.3.4 Pre-Treatment only Integration as Reference Performance Data

Integration of microarray data of pretreatment only data in the adjuvant therapy setting is an established practice with early investigations identifying and accounting for the multiplicative distortions introduced in the routine processing of microarray data.[240] Subsequent work by Turnbull et al. showed, for example, the cross platform normalisation of Affy and Illumina data using XPN and *ComBat*,[216] or Greene et al. for a further instance of XPN's application to the simultaneous evaluation of microarray and RNAseq data.[226] Removing batch effects with techniques like *ComBat* have been shown to facilitate cross platform expression comparison.[241] The pretreatment samples were integrated using quantile normalisation to standardise expression levels and ComBat with platform as the "Batch" effect to be corrected. Performance of *ComBat* in this setting will be quantified to provide a reference for the amount of distortion added by the inclusion of the on-treatment samples.

4.3.5 Uncorrected Integration Methods

Uncorrected approaches to dataset integration were performed with subset (retaining only the expression features common to all datasets) and complete (retaining all features across all datasets) joins of data. Missing genes post join were filled with added interpolation of missing features to preserve dimensionality (i.e. keep full gene annotation) across datasets, this was performed with K-Nearest-Neighbor imputation, Figure 4.3.2 highlights how these values are calculated, where the K nearest neighbors are located and the missing values are filled in with the mean value from the N neighbors. Quantile normalisation[223] was used to retain the ranking of gene expression on a per dataset basis, while converging the expression distributions. Multiple tests were then used to score the performance of these methods and used as a second reference of performance for the *ComBat* assisted pre- and on-treatment sample integration.

As previously stated, complete observational and subset joins will result in dramatically different dimensions for the product dataframe. While the number of patients will not differ between the two methods, the feature list (rows) of the integrated data will reflect the difference in the mechanics of each join. In the case of this translational data, the features are genes. The resulting size of the gene lists are plotted below as venn diagrams. The final dimensions for each join are: Full outer join, 1122 samples with 21075 gene annotations, inner join 1122 samples with 4786 gene annotations.



	Samples with NAs			Imputed Values		
	Sample 1	Sample 2	Sample 3	Sample 1	Sample 2	Sample 3
Gene 1	8.52	NA	11.23	8.52	11.23	**9.88**
Gene 2	9.12	7.47	8.44	9.12	7.47	8.44
Gene 3	7.98	NA	9.20	7.98	9.20	**8.59**

Figure 4.3.2: Example Imputed Values from Continuous Gene Expression Data. KNN imputation works by finding the K nearest neighbours, averaging the values from those neighbors and supplying them as the missing values. In this example Gene 1 and Gene 3 are missing values for Sample 2, indicated by the blue arrow. KNN in this example is looking for the 2 nearest neighbours, Samples 1/3 indicated by the green arrows. It will then average the values of the missing genes in these samples to impute the missing values for Sample 2. The blue dots represent other samples which are more distant than Sample 1/3 and are not nearest neighbors in this case.

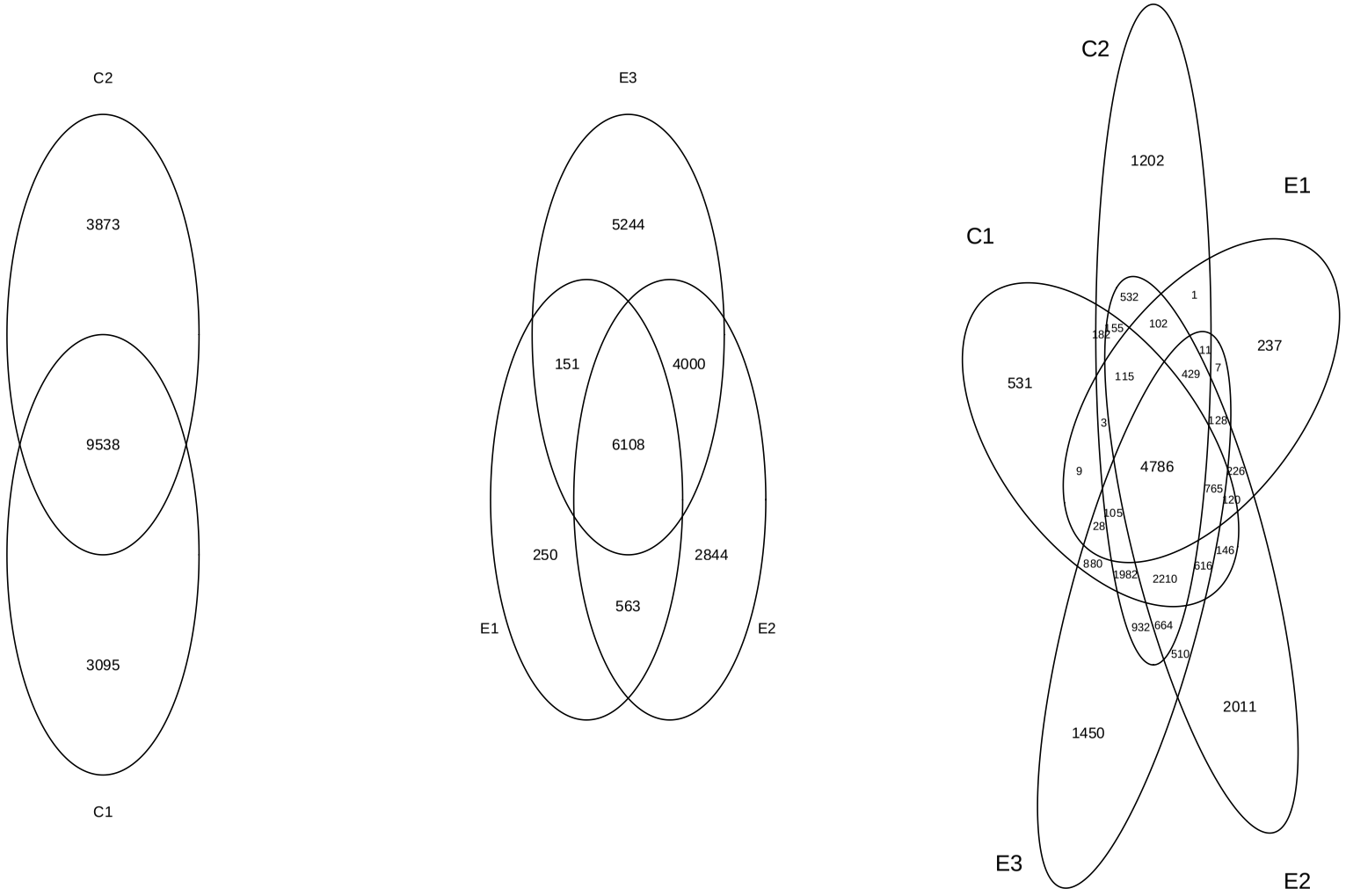


Figure 4.3.3: Comparison of The Numbers of Commonly Represented Genes When, 2, 3 or 5 Datasets Are Combined. The 2 lobed venn-diagram is the overlap of the two chemotherapy cohorts, the 3 lobed venn-diagram represents the overlap of the 3 endocrine therapy cohorts, and the 5 way venn-diagram is the overlap of the features present in all five datasets.

4.3.6 Reproducible *ComBat* Workflows and Integration Testing

Integration of the pre- and on-treatment data was performed using the pre-selected covariates of analysis; Platform (P), Therapy (TT), Time (TS), and Subtype (ST). These covariates were chosen as they were the most uniformly represented across the datasets. Every sample had at least one matched on treatment or surgical biopsy sample (TS) and the hypothesis that the on-treatment samples are markedly different to pretreatment had been previously examined and determined to be true in chapter 2 which details the NEO study, and chapter 3, which details the primary and node trials. Every sample came from a distinct platform (P) and cross platform normalisations are a normal method of integrating numerical expression data. Therapy was equally well annotated as every sample either came from a Chemotherapy or Endocrine dataset and the difference between treatment types was believed to be evident in the resultant expression level changes. Lastly subtype can be derived for every sample is an important metric for monitoring stability of the integration efficacy as large changes would be indicative of large changes to the underlying relative expression levels. Clinical/pathological variables were immediately excluded because they were sparsely included in the sample annotation files and would have detrimentally reduced the sample size. Purrr[242] was used to functionalise the testing process and perform the integration on all possible combinations of integration covariates. Purrr is an R library for applying functions, or lists of functions to lists of objects and helped to keep the analysis uniform across the datasets as they could be analysed completely in parallel. Combination of the covariates was not included in this analysis as the ballooning effect of the composite covariates would have quickly made this a functionally different task. The effect of the above covariates was not know either and the combination was a logical extension if certain levels of performance of integration were met.

As in previous results chapters, the primary method of reporting significance between two groups, or measuring the significance of the difference present between to effect populations in this study is through a combination of parametric and non-parametric tests. When measuring the effect between groups of unpaired samples standard T-tests were performed. The five basic assumptions were always tested: The values are on a continuous scale (expression values), the samples are to the best of my knowledge representative of the population as a whole as with the increased sample size a greater representative cross section of the general population is represented than would exist in any one data set in isolation. Graphically the data conforms to a normal bell shaped distribution

after processing as can be seen in Figure 4.2.2, with the exception of the second on-treatment samples of dataset C1. The fourth criteria of a reasonably large sample size is satisfied as the other assumptions hold and there are sufficient samples to calculate the test statistic. Lastly, standard deviations on both side of the mean are approximately even suggesting homogeneity of variance. Non-parametrically a Wilcoxon T-test was used when measuring the differences in expression of paired samples. Assumptions here were even easier to satisfy as every sample is inherently paired and they all come from the same population, additionally, we are looking at looking at within pair differences.

4.3.7 Patient Matched Concordance and Correlations as Integration Metrics

Patient-matched correlation, heatmaps with hierarchical clustering, principal component analysis, intrinsic subtype concordance, differential gene expression, machine learning classification and pre- to on-treatment proliferation fold change values were compared to establish discordance between the reference data and each iterative integration attempt. Concordance and discordance will be measured by the amount of overlap present for a specific discrete metric after an integration step and are inversely proportional. Discordance among subtypes for instance would indicate that the predicted subtype for a sample has changed post integration, and concordance would indicate that it has stayed the same. Heatmaps were generated for every *ComBat* integration attempt using the R package `pheatmap`[243] with hierarchical sample clustering and per patient annotation of the different biological and systematic features using Euclidean distance, complete observational linkages and no robustness testing. Initial subtypes and risk scores for IC10,[154–156] Pam50,[156] scmod1[157] and scmod2,[158] ssp2003[154] and ssp2006,[155] MammaPrint[159] and rorS[156] continual and categorical risk values were calculated. Machine learning methods for the classification of cancers[244] and the prediction of recurrence of breast cancer[245] is well established. In particular, random forest has been shown to be a robust algorithm for breast cancer risk prediction.[245] This is true of both RF classification and regression models that return categorical risk or continuous risk respectively. There are other methods that are established for this regression/classification problem from linear regression models to neural networks. It is the findings of previous work in this thesis that RF generalises well to data of this scale and noise and due to the inherently democratic method of RF avoids some of the over-fitting of simpler regression models, but can be reliably trained unlike more sophisticated deep learning methods. Follow up analysis of cosine similarity was performed using scikit-learn[149] to compare the inter-patient similarity pre- and post inte-

gration.

4.4 Results | Unintegrated Data Meta-analysis

In order to understand how each attempt at integration affected the underlying biology of the patient samples baseline measurements were required as a reference. The following sections highlight the important diagnostic results from these baseline tests to help make future comparisons.

4.4.1 Subtype Composition of Unintegrated Datasets

The composition of the unintegrated data is presented in Figure 4.4.1, Page 89 as pairwise risk score comparisons, annotated by the color of the initial intrinsic subtype calculation. Four important pieces of information are contained in this diagram. First, there are relatively more treatment matched patient samples for the aromatase inhibitor treated datasets. Second, there is a trend in the data for a reduction in the calculated risk of the samples on treatment compared to pre, this is evidenced by the position of the sample lower on the Y axis (On-Treatment) relative to their corresponding X axis (Pre-Treatment). Third, there is an over representation of Basal subtypes in the chemotherapy cohorts compared to the aromatase inhibitor treated cohorts and lastly, the most significant reductions in risk are those of the on-treatment Luminal B samples in the aromatase inhibitor datasets (wilcoxon $p < 0.05$ for mean reduction). Comparisons of all 8 classifiers (pam50, ic10, ssp2003|6, smcgene, mammaprint, rorS, scmod1/2) will be made, with results presented as concordance between the pre and post integration values on a per patient basis.

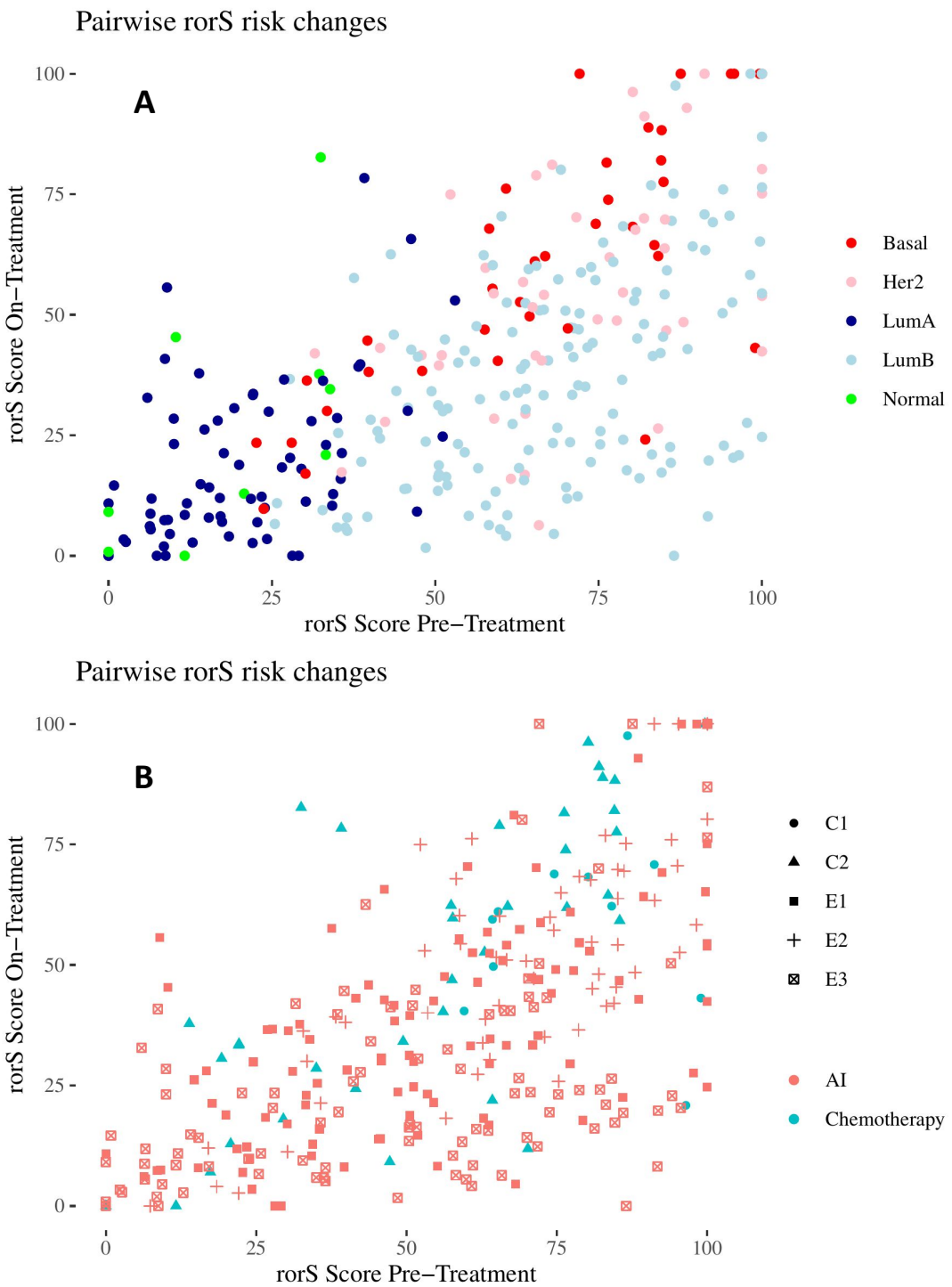


Figure 4.4.1: Independent Analysis of the Subtype Composition Across Multiple Datasets. This figure illustrates the pre integration patterns of predicted risk change with treatment. The X and Y coordinates are the TP (Pre-Treatment) rorS risk and T-On (2 weeks on treatment) rorS risk, respectively. Each sample is coloured by their Pam50 assignment at TP to illustrate subtype specification changes. A) shows the samples coloured by the Pam50 subtype, B) shows the samples coloured by treatement received and given a shape based on the dataset of origin.

4.4.2 Continuous and Categorical Risk Classification Across Multiple Independent Datasets

The relative risk and relapse scores for each of the five datasets were examined using MammaPrint and PAM50 rorS. These scores were calculated to identify differences in the presentation of the datasets and to act as a reference for the changes in the apparent risk associated with each dataset after each method of integration. The median risk score and the tally of categorical risks classifications for each dataset according to both rorS and MammaPrint is presented in the following Table 4.3, Page 91. The input for this geneFu method are expression matrices like with the classification tests. No specific method of normalisation is required for any of the features estimated here. The output for these tests are two fold, first a continuous risk score from -1 (lowest risk) to 1 (highest risk) for MammaPrint and between 0 (lowest risk) and 1 (highest risk for rorS, second a categorical binning of patients into classes of 0 (low risk) or 1 (high risk) for MammaPrint, and Low, Medium and High for rorS. Continuous risk scores are presented as means with 95% confidence intervals. Clinically it is interesting to see that the chemotherapy prescribed cohorts are of higher average risk across both tests, with a greater proportion of categorical high risk samples. The two chemotherapy cohorts were comprised of patients with a worse general prognosis and average higher histological grade. Deviation from these values, continuous or categorical will inform us on the affects of integration to the prognostic appearance of the patient samples. E2 is aberrant here for an endocrine dataset and appears much higher risk than the other. This is, at this point, difficult to explain as the data has been normalised and tests with the same parameters. The composition of the cohort was known to be of lower clinical risk as well, suggesting some instability, possibly, with this method.

Dataset	MammaPrint Risk	MammaPrint Score	rorS Risk	rorS Score
E1	0(175), 1(187)	-0.29(-0.309 to -0.271)	L(141), I(111), H(110)	36.3(34.8 to 37.2)
E2	0(0), 1(176)	-0.04(0.0128 to 0.0672)	L(35), I(49), H(92)	54.6(53.1 to 56.1)
E3	0(191), 1(27)	-0.45(-0.476 to -0.424)	L(116), I(46), H(56)	26.1(24.7 to 27.5)
C1	0(29), 1(66)	-0.23(-0.267 to -0.193)	L(35), I(24), H(36)	40.8(38.4 to 43.2)
C2	0(56), 1(192)	-0.004(-0.0269 to 0.0189)	L(62), I(55), H(113)	55.8(52.7 to 58.9)

Table 4.3: Continuous and Categorical Risk Assessment Across Multiple Independent Datasets. This table presents the estimated risk associated with each data set. The chemotherapy datasets were of generally higher average risk and had more categorical high risk patients. The columns of this table represent the MammaPrint categorical risk, the MammaPrint continuous score, the rorS categorical risk and the rorS continuous score. 0 represents low risk, 1 represents high risk, for rorS L represents low risk, I intermediate and H high.

4.4.3 Pre-integration PCA Visualisation Data

PCA of the unintegrated data is presented Figure 4.4.2, Page 93. The principal components are the eigenvectors that describe the variance in the data, i.e., the X and Y displacement of each point on a 2D axis. These eigenvectors are made up of the contribution of many features, thus the position in space of one sample relative to another indicates the similarity of those two samples. Samples from the same cohort or the same patient would be expected to cluster near each other, for example. The effect of a platform batch effect is clearly visible in the PCA representation. As expected with highly dimensional data the amount of explained variance is low, at 17.2% and 12.1% in PC1/2 respectively.

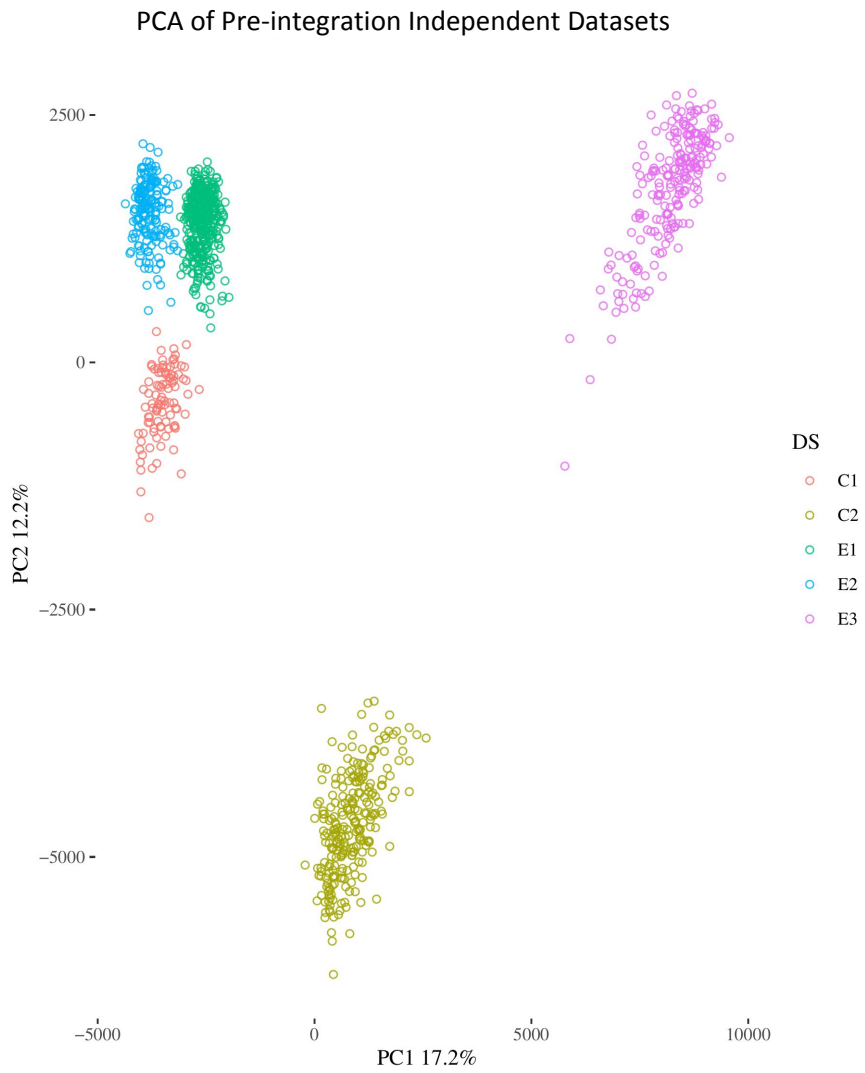


Figure 4.4.2: PCA Diagram of the Unintegrated Data Illustrates the Strong Presence of Platform Associated Batch. This PCA visualisation represents the five datasets, colored by the cohort of origin. It is clear from this initial analysis that there is distinct seperation of the data on the basis of batch.

4.4.4 Principal Components Analysis of Independent Unintegrated Data

Some of the genes that comprise PC1 and PC2 for each data set pre-integration are presented in Table 4.4, Page 94. These gene lists will be use for later comparison of Principal Component outputs to enumerate the degree of change caused by each integration attempt.

Dataset	Genes in PC 1	Genes in PC 2
E1	HSPB6, MYH11, PLIN1, RBP4, EGR1	AGR3, AGR2, MUCL1, SCGB2A2, TFF3
E2	TTK, DUS1L, DHX33, CDCA8, C10orf54	TFF3, KRT19, KIAA1324, FOXA1, RARRES3
E3	LYPLA1, ARMT1, TMEM106B, CAPZA2, VPS4B	DPT, FABP4, FOS, ACKR1, C1S
C1	RPL41, RP5, RPS23, EEF1A1, RPL13A	COX6C, CPB1, PIP, CARTPT, SLC39A6
C2	RGS1, MNDA, EHF, SPARCL1, MYSM1	NKAIN1, SYT13, DSCR6, PDZK1, GRIK3

Table 4.4: Contents of PC1 and PC2 for Future Comparison. Top five contributors of the first and second principal components for each dataset are listed above, these are primarily a resource for comparing the change in the variance as a product of integration. The biological relevance of these gene lists has not been examined as the method for identifying them is not to elucidate biological processes or pathways but as a comparative metric.

4.4.5 Analysis of underlying Data Structure Through Similarity Scores

Additional analysis of the PC results have expanded this study to include the differences between different patient pairs pretreatment and post-treatment, the results of this test can be found in Table 4.5, Page 95. The purpose of this analysis was to determine if the structure of the data is fundamentally different pre and post integration and to measure this cosine similarity was used between all samples and all other samples and then averaged to create a single point for comparison. This metric is calculated as the inner product between two normalized vectors, which is approximate to the cosine of the angle between them. This was used in this instance as an alternate metric to correlation as it can calculate the distance between two vectors, samples, taking into account the sample in n-dimensional space. This means we are in effect calculating the movement away from the starting position post integration for every sample relative to every other sample. It is a supplementary method to the correlation calculations. These results suggest that the integration process even at this level of complexity has a systematic difference not only on the presentation of the individual samples but on the data as a whole. Comparison of key biological genes or housekeeping genes is being omitted as it is clear from the overwhelming differences in average similarity between the integration steps that any comparison

between these genes would yield similar differences.

	Inter-Patient Avg. Similarity (Pre-Integration)	Inter-Patient Avg. Similarity (Post-Integration)
C1	0.43(0.389 to 0.471)	0.67(0.619 to 0.721)
C2	0.55(0.491 to 0.609)	0.71(0.636 to 0.784)
E1	0.51(0.454 to 0.566)	0.66(0.59 to 0.73)
E2	0.84(0.78 to 0.9)	0.82(0.719 to 0.921)
E3	0.61(0.56 to 0.66)	0.73(0.668 to 0.792)

Table 4.5: Table of the Average Cosine Similarity for Each Dataset This table shows the average cosine similarity between all samples for each dataset pre- and post-integration. The pre-integration scores are all lower indicating less similarity before treatment, this is possibly representative of greater variation in the relative gene expression that appears to be some what homogenised post-integration. While speculation on the rationale behind the value change is ultimately moot, the salient point is that the underlying data structure has fundamentally been altered. It is worth noting that as usual dataset E2 appears as some what of an outlier in this analysis.

4.5 Results | Correlation Analysis

Pearson correlation of matched patient samples was calculated using the base R `cor` function to establish high level similarity between samples pre and post integration. This was used as a metric to gauge the similarity or dissimilarity of the full transcriptomic profile of each patient sample.

4.5.1 Pretreatment Samples Corrected with Platform ComBat Batch Correction

Correlation of individual samples is presented in Figure 4.5.1, Page 96 showing how patients with differing similarity appear for reference post integration. Matched patient correlation on a per-dataset level is presented in Figure 4.5.2, Page 97 to visualise the correlation of pretreatment samples pre and post-integration. Correlation values vary between the datasets, but are consistently high (75% >). It is difficult to establish a firm reference metric for acceptable correlation values, but this level of correlation is strongly indicative of a trend.

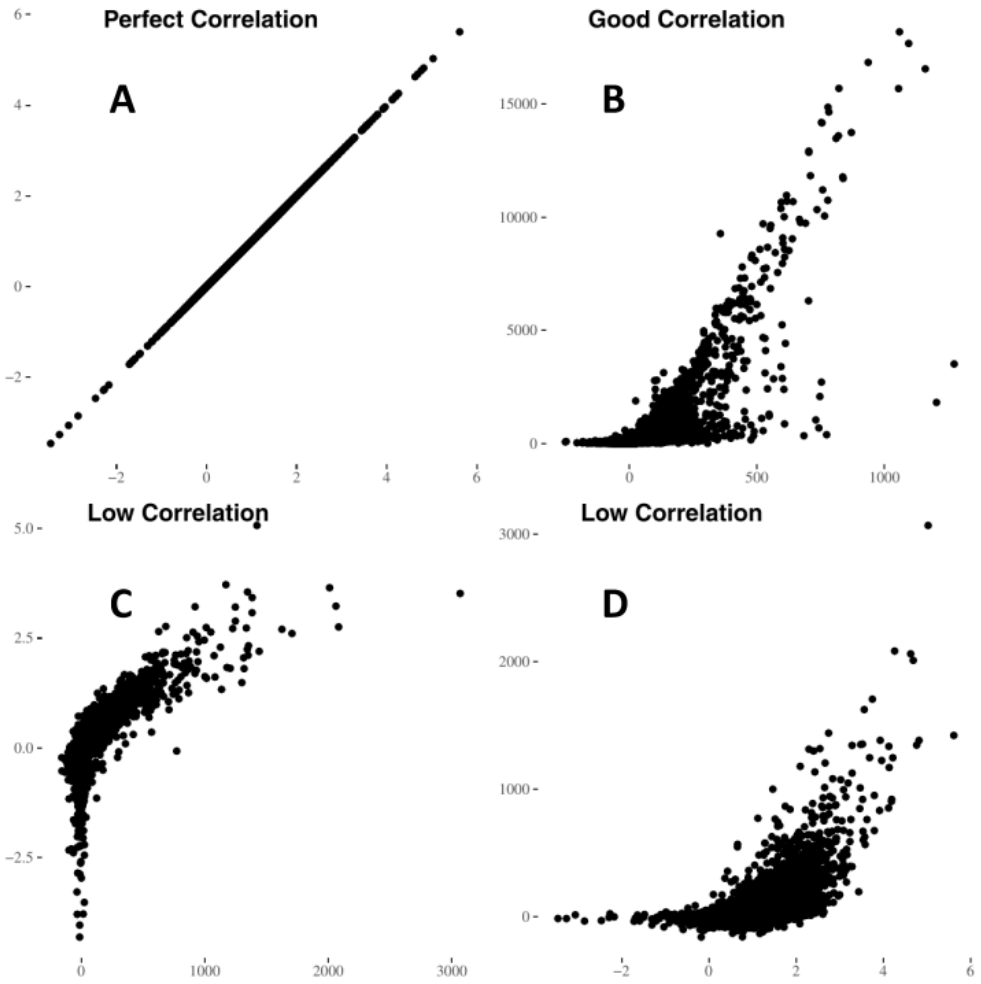


Figure 4.5.1: Example Correlation Diagrams of Purely Synthetic Data The data displayed in these correlation diagrams are purely artifical and are not representative of specific patients, samples or studies. These are illustrative only. A) Shows a hypotehtical correlation of 1, indicating no change in the pre and post integration samples. B) A correlation of 0.88, indicating a low level of distortion between the pre- and post-integration sample. C) and D) are samples with a correlation of 0.22 and 0.24 respectively, which would be a very low correlation suggesting a large amount of integration distortion.

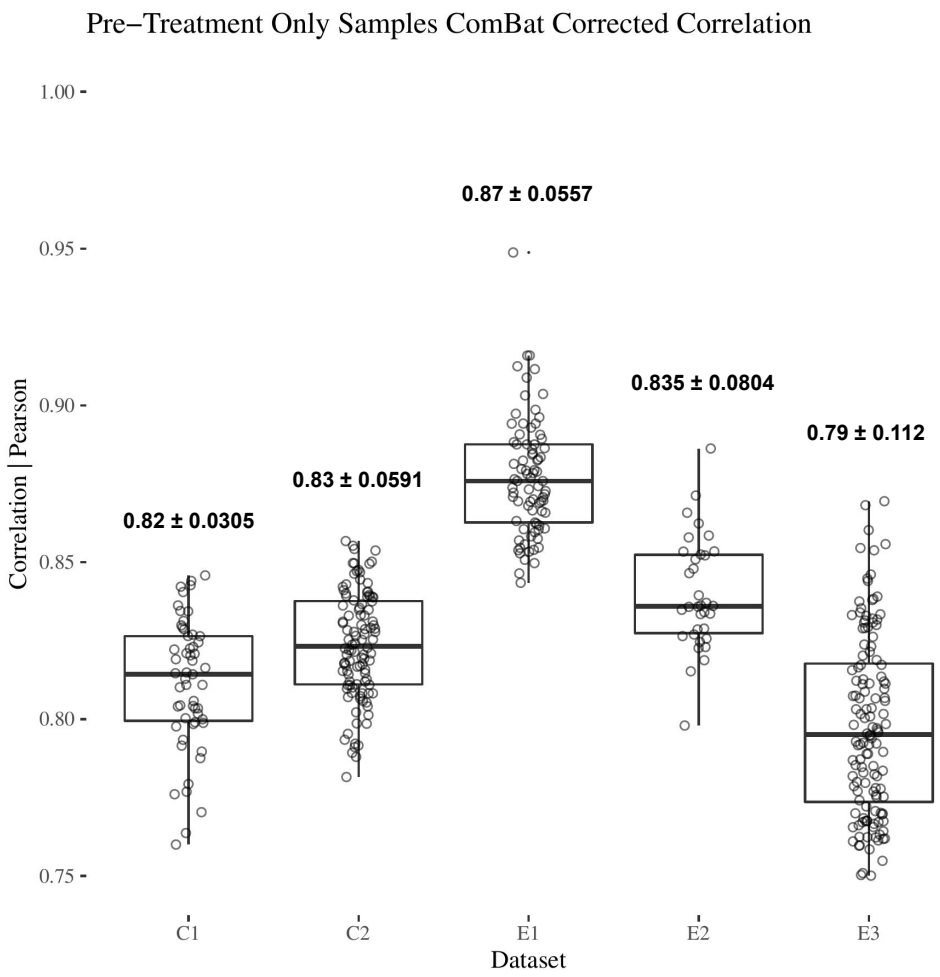


Figure 4.5.2: Correlation of Pretreatment only Samples Using Combat with Platform as a Covariate Correlation values on a per sample level per dataset are shown in these box plots and show the changes that are caused by the integration on a superficial and unsupervised level. There is a high correlation of the data pre to post integration when considering only the pretreatment samples, with a range of mean from 0.72-0.87 (represented by the middle line of the boxplot) and 95% confidence intervals are the terminal ends of the boxplot "Whiskers") and presented in the figure.

4.5.2 Uncorrected Combination of Sequentially Sampled Pre- and On-treatment Datasets

Figure 4.5.3 Page 98 shows the per dataset correlation for the uncorrected feature joining methods. Both methods showed similar levels of average correlation to the pretreatment only. E2 is the most tightly correlated for both methods, and with the exception of E3, there is little statistical difference in the average correlation of the data (all pvalues <0.05 using Wilcoxon t-test).

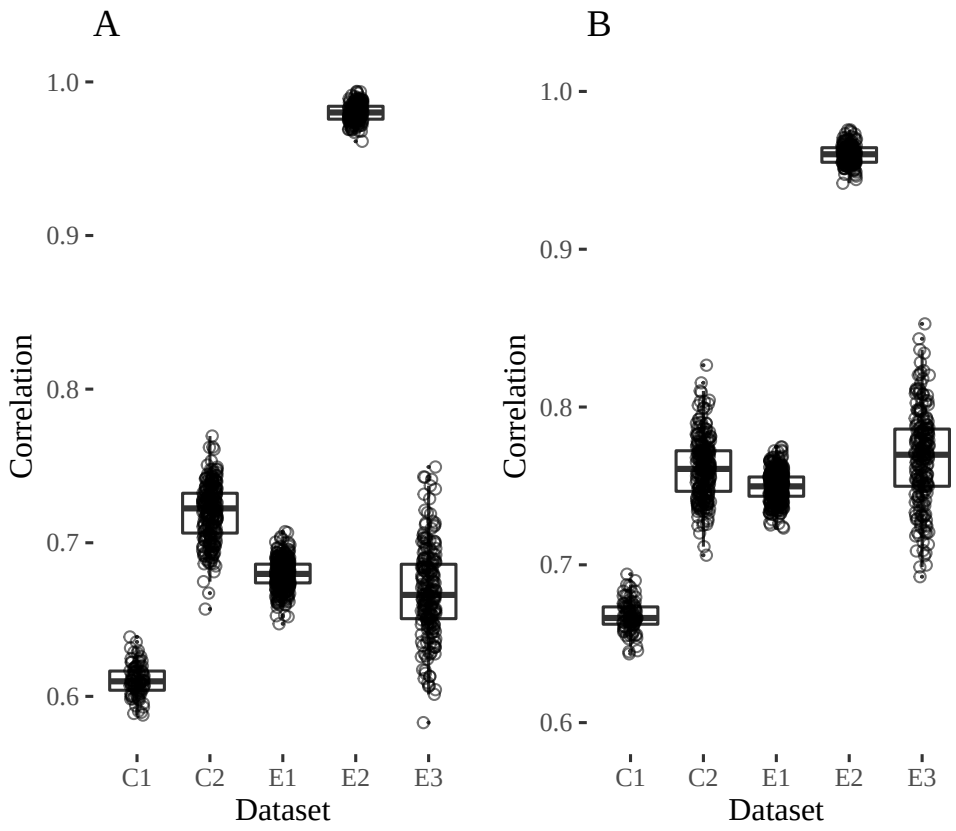


Figure 4.5.3: Matched Patient Sample Correlation for Uncorrected Feature Joining Methods. Correlation of patients to their matched samples post combination by complete (left) and subset joins (right) show the resultant per-patient correlation represent on a per-dataset level. This shows a clear drop in correlation post combination, with the exception of E2. Mean values are indicated by the midline, with upper and lower quartiles represented by the bounds of the box and the 95% confidence intervals as the terminal ends of the whisker plots.

4.5.3 *ComBat* Integration of Sequential Samples with Multiple Covariates

It is clear from Figure 4.5.4, Page 99 that there is a range of correlations. The E2 cohort is of significantly higher correlation with its own matched preintegration samples than the other cohorts ($p < 0.05$ Wilcoxon T-test). Summary table of all correlation comparisons for QN|Inner/Outer-join and all four ComBat methods are available in Table 4.6, Page 100

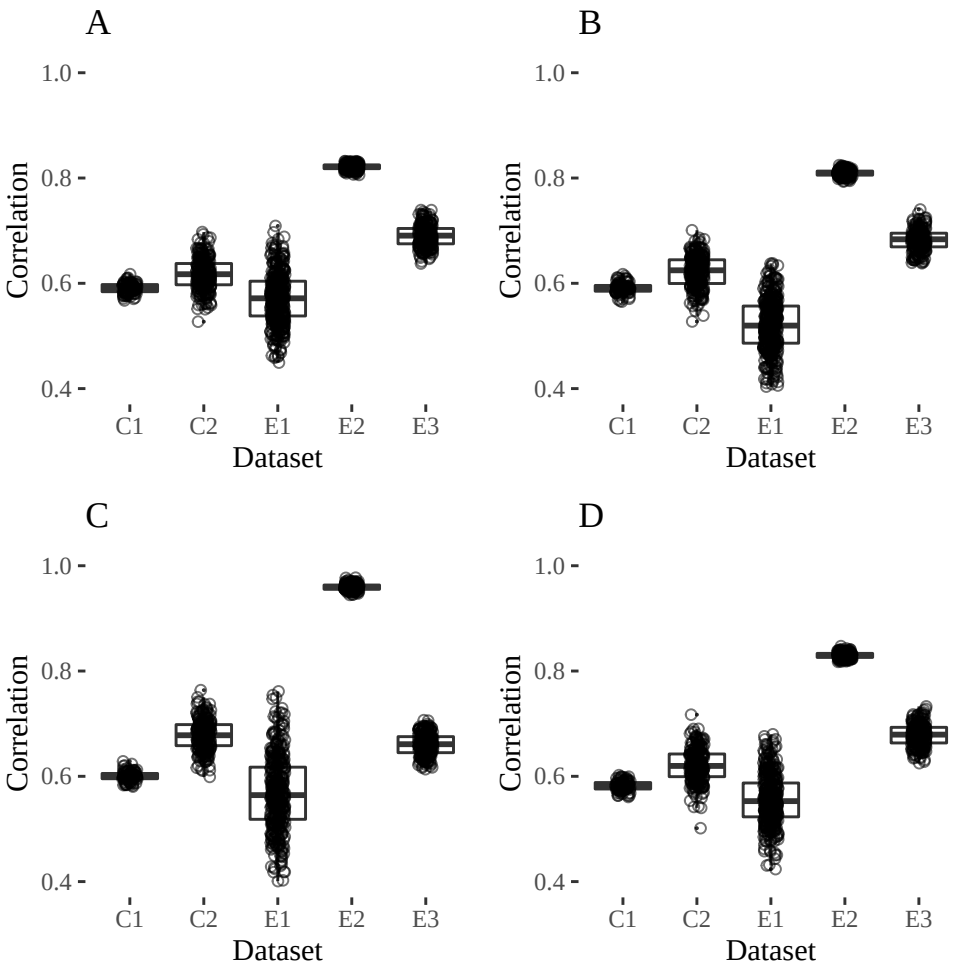


Figure 4.5.4: Correlation Pre- and Post-Integration are Lower for All ComBat Covariates. Correlation is stable across the four integration methods (Platform (A), Time (B), Therapy (C) and subtype (D) and comparatively low when compared to either uncorrected methods, or the pretreatment only sample integration.

	QN: Inner Join	QN: Outer Join	Platform	Treatment Status	Treatment Type	Subtype
C1	0.62(0.59-0.64)	0.66(0.64-0.69)	0.59(0.58-0.60)	0.59(0.58-0.60)	0.6(0.59-0.61)	0.59(0.58-0.60)
C2	0.73(0.68-0.77)	0.76(0.71-0.81)	0.62(0.56-0.68)	0.62(0.56-0.68)	0.63(0.57-0.69)	0.62(0.56-0.67)
E1	0.68(0.66-0.70)	0.75(0.74-0.76)	0.58(0.48-0.69)	0.53(0.41-0.63)	0.58(0.41-0.7)	0.58(0.46-0.66)
E2	0.98(0.97-0.99)	0.96(0.95-0.97)	0.84(0.82-0.85)	0.83(0.81-0.84)	0.93(0.91-0.94)	0.83(0.82-0.84)
E3	0.66(0.63-0.74)	0.78(0.70-0.85)	0.68(0.63-0.72)	0.67(0.62-0.71)	0.66(0.61-0.7)	0.67(0.64-0.73)

Table 4.6: Mean Correlation and CI for Each Method for Each Dataset This table contains the mean correlation values and the range for the 95% CI for each method for each dataset.

4.6 Results | PCA Analysis

PCA analysis reveals how the variance within each dataset is affected by integration, at least to the extent of the first and second principal components. Diagrams of the first and second principal components were created to visualise the change in variance and evaluate the performance of the integration methods. Additionally, gene lists that describe the first and second principal components of the dataset where compared pre and post integration to examine the similarity in explained variance. This study wants to acknowledge that when using such dimensionality reducing techniques such as PCA there are other techniques, like UMAP or t-SNE, that may be equally viable. Looking at arbitrary variance cut offs and comparing the Principal Component contents, or the matrix differences of reduced space would be two possible avenues. In addition not looking at the transformed rotation values but looking at the loading values, the weights in the model that multiply the original scores to create variance unites, you can also compare the way in which the data is changed. This was not followed further as the onus of evidence was suggesting that any further analysis would only yield concurrent results and was a duplication of effort.

4.6.1 Pretreatment Samples with Platform ComBat Batch Correction

The 2D representation of the first two principal components can be seen in Figure 4.6.1, 101. This was produced using only the multiplicative variable of Platform, and making non-parametric estimation to the prior distributions. This attempt at integration is visually conforming to the known *ComBat* output of "successful" integrations of microarray and expression data.

Pretreatment only sample integration resulted in an average retention of PC1 gene lists of 53.6% and PC2 of 32.12%, Table 4.7, Page 102. This result indicates that in the two eigenvectors explaining the variance of the data, there is at best a similarity of about a half and a third of the same genes.

Figure 4.6.1: Integration of Pretreatment Only Samples Using QN and ComBat Integration of the pretreatment only samples clearly shows normal and expected integration using the literature established methods of integration for transcriptomic data. A) The pre-integration uncorrected PCA results represented by Figure 4.4.2, from Page 93, B) the pretreatment only integrated samples.

	Pre-Treatment Only		Complete		Subset		Platform		Time		Therapy		Subtype	
	PC1	PC2	PC1	PC2	PC1	PC2	PC1	PC2	PC1	PC2	PC1	PC2	PC1	PC2
C1	52.2%	32.4%	11.8%	0%	0.2%	1%	0.2	0%	11.6%	4%	0.4%	10.4%	10.8%	0.4%
C2	61.6%	19.8%	68%	5%	0%	23.6%	2%	0.4%	60.2%	14%	0%	20.8%	65.8%	18.6%
E1	42.7%	14.1%	33.6%	7%	28.6%	8.8%	0.8	0%	22.6%	1%	23.4%	47.2%	23.4%	1.2%
E2	82%	21%	62.8%	14%	18.4%	26.6%	1.6	0%	15.2%	11.8%	59.4%	5.2%	27.4%	14.2%
E3	29.5%	73.3%	18.4%	26.6%	1%	0%	0.6	1%	20.2%	37.8%	14.8%	35%	16.4%	37.4%

Table 4.7: Principal Component Overlap Between Reference Data and Each Integration Method. Each row contains the percentage of genes present in the first and second principal component of the post integration data present in the preintegration data for each combination method. Analysis of principal components was capped at 2 components as there was such as the variance loss past PC2 was extensive, and was not adding anything to the analysis.

4.6.2 Uncorrected Combination of Sequentially Sampled Datasets

Principal component analysis of these two methods is shown in Figure 4.6.2, 104. It is clear from the separation of the groups that the batch effect is still present and that normalisation alone has not reduced the intergroup differences at all. The scale of the variational distance is different between the two gene lists in PC1, this is likely due to the feature size of the data but despite this, the pattern of clustering is comparable. The percentage of overlap of the first and second principal components is presented in Table 4.7, Page 102. There is an average PC1 overlap of 38.92% and an average PC2 overlap of 10.52%, comparably lower than the pretreatment only integration values.

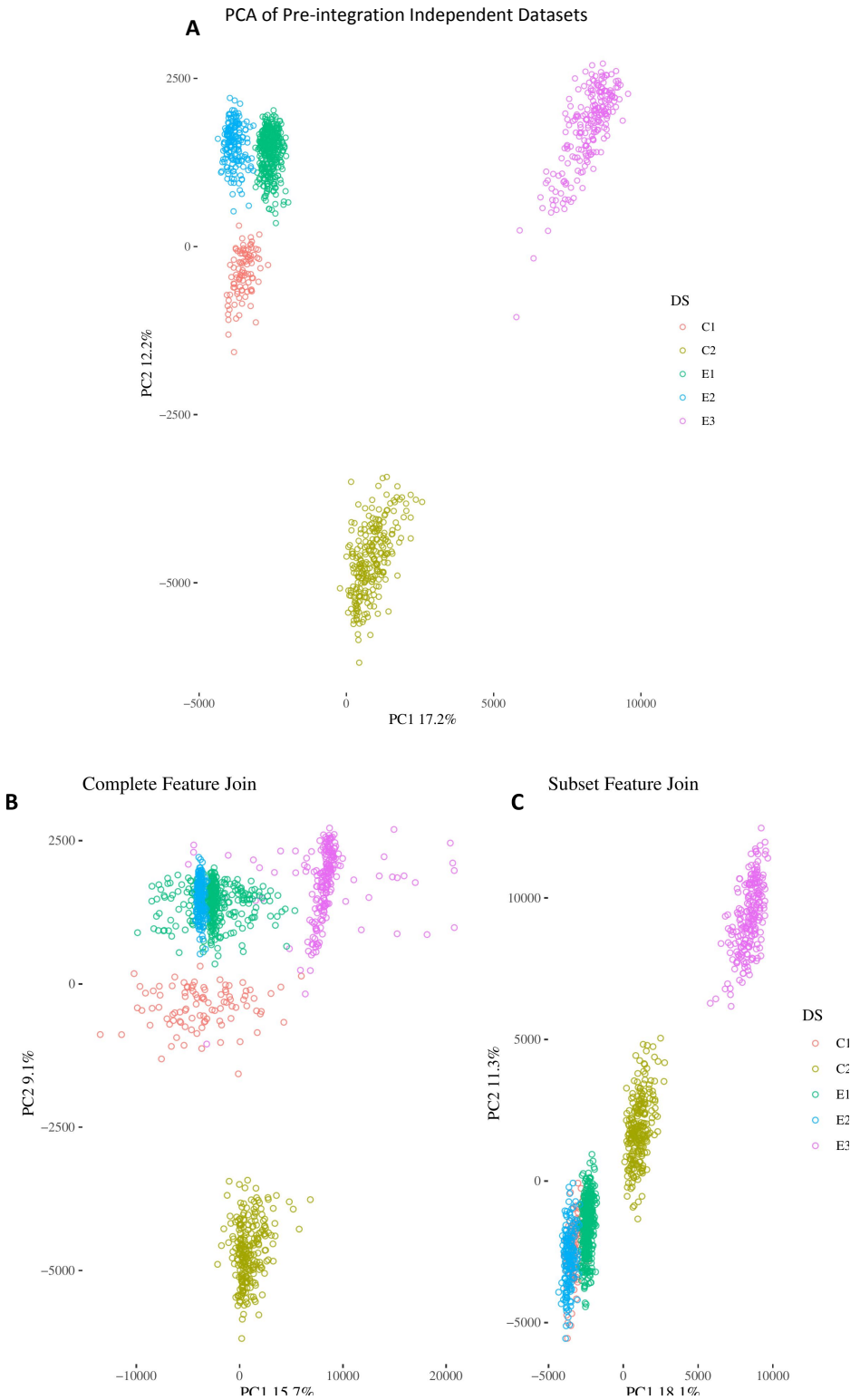


Figure 4.6.2: Uncorrected Combination Methods Do Not Improve Dataset Integration Changing the contents of the gene lists for the resultant dataset changes the variance contributions of each principal component. Here the variance is visually represented in a PCA diagram, and the difference in inter-dataset distance are clearly apparent between the two methods. A) Represents the uncorected data visualised by Figure 4.4.2, from Page 93, B) the joined data that results from a full outer join, C) the joined data from only an inner join all performed with QN.

4.6.3 *ComBat* Integration of Sequential Samples with Multiple Covariates

PCA of the post integration data reveals significantly less visual overlapping of the datasets than the pre-integration only or uncorrected methods, Figure 4.6.3, Page 106. Additionally, with increased feature count for the additive covariate there is increased fragmentation of the resultant PCA plot. With respect to the pretreatment only, there is a clear loss of structure and batch, and compared to the uncorrected joins there is an increase in the number of visible batches. Table 4.7, Page 102 contains the overlap values of the first and second principal component, there is a mean overlap of PC1 and PC2 of 1.04 and 0.28, 25.96 and 13.72, 19.6 and 23.72, 28.76 and 14.36 for the covariates of platform, treatment time, treatment type and subtype respectively.

Theoretically, correlation between principal components should be low as they are calculated from non-correlated, orthogonal eigenvectors with unique contributions from each gene. As can be seen in Table 4.8, Page 105, this trend is true for all datasets except for E2. The collinearity seen in E2 shows that the features that define PC1 are related to the features in PC2, indicating that variance in one plane effects variance in another.

Dataset	Platform	Time	Therapy	Subtype
C1	12.9%	13%	15.1%	15.0%
C2	6.7%	9.2%	13.1%	11.6%
E1	9.1%	8.7%	11.8%	11.2%
E2	87.2%	89.6%	86.8%	89.7%
E3	14.1%	12.2%	15.3%	11.1%

Table 4.8: Measurement of Collinearity Between PC1 and PC2 This table contains the values of collinearity between the first and second principal components for the five datasets across the four batch aware integration methods. Principal components should in theory be orthogonal and thus have no linearity, this would be a score of 0%, full collinearity would be 100% and suggest that the vectors that represent PC1/2 are not collinear at all.

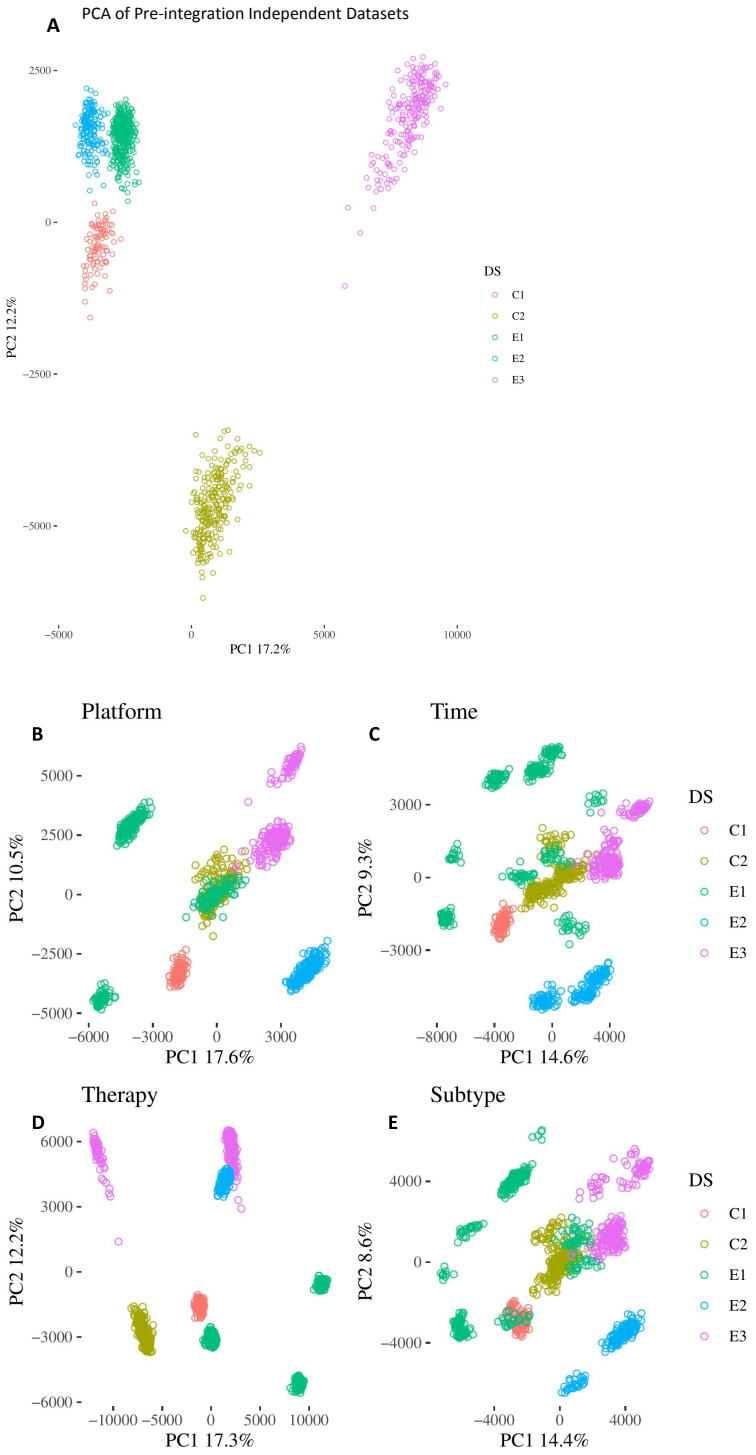


Figure 4.6.3: PCA Plot Comparisons of Four Integration Methods. Comparing PCA plots of different integration methods Platform, Time, Therapy, and Subtype reveal a decomposition of the original batches. A) The reference PCA plots of the pre integrated data represented by Figure 4.4.2, from Page 93. B) Shows the Platform aware integration, C) Treatment Time, D) The therapy used, E) The Pam50 subtype. These results show many different structures distinct to the covariate of integration.

4.7 Results | Heatmap Visualisations

In concert with the PCA diagrams, heatmaps can highlight groupings based on the clinical and pathological features available for analysis. Heatmaps of the pre-treatment only sample and uncorrected integration methods were omitted due to the PCA diagram results. Heatmaps were generated for every *ComBat* integration attempt using the R package `pheatmap`[243] with hierarchical sample clustering using euclidean distance and complete observation linkages and per patient annotation of the different biological and systematic features. These were used for visual inspection of the effectiveness of *ComBat* and to estimate the level of integration based on the presence or absence of noticeable groupings of the annotated features. In addition Silhouette coefficients were calculated for each clustering analysis in order to enumerate and analyse the cluster robustness.

4.7.1 *ComBat* Integration of Sequential Samples with Multiple Covariates

Post integration heatmaps of the combined datasets were plotted to visualise the representation of the biological and systematic annotation features among the resultant batches. As Figure 4.7.1, Page 108 shows, there is no significant clustering of any of the annotation features relevant to this analysis. There is some mixing of C2 (Blue) and E1 (green) upon integration with covariates of Platform, treatment type and subtype, however the datasets are not integrated with respects to the treatment received, nor the patients by response or treatment time. Fundamentally there appears to be present batch effects in the post integration data that are not corrected by the integration methods. As a measure of the mixing of the samples the two closest neighbors were calculated for each sample. For ninety eight percent of the samples the nearest neighbor originated from the same dataset. This suggests a low level of integration as the dataset of origin was the most prominent feature of the post integration space. In addition, the Silhouette Coefficient for each each data set with each integration method was above 0.5, and with a mean value of 0.881, strongly suggesting that the clustering of each dataset is tight and distinct from the other datasets post integration Table 4.9, Page 109.

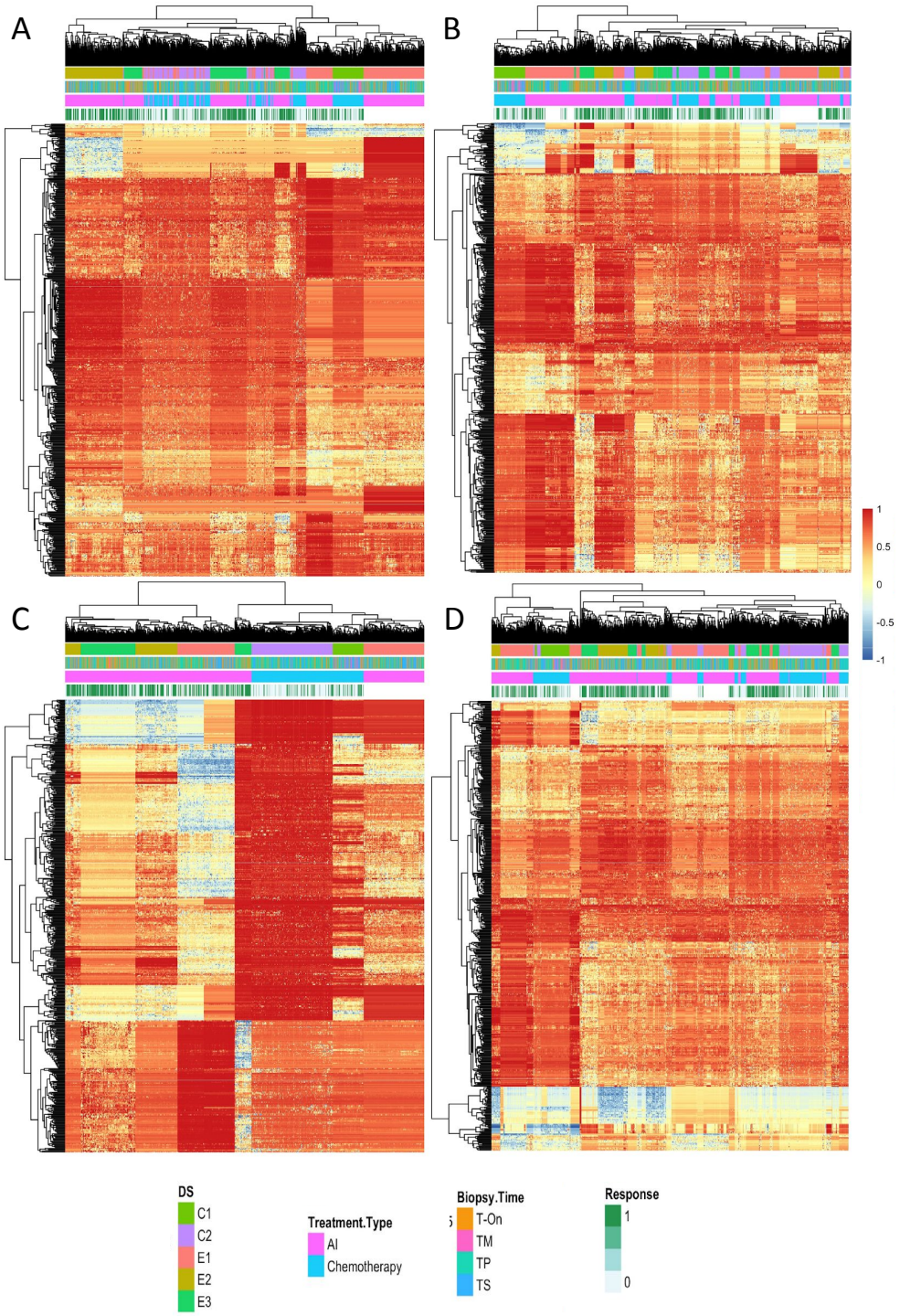


Figure 4.7.1: Heatmaps and Hierarchical Clustering of Post Integration Datasets. Heatmaps (HM) and clustering can help to identify patterns, or visually confirm results of analysis. Here, the patterns of non-integration are clear from the distinct expression patterns per dataset, and the visible lack of integration of the annotaion features. HM A shows the clustering an expression patterns for Platform integration, B for Time, C for Therapy and D for Subtyping integration.

	Silhouette Coefficient				
	E1	E2	E3	C1	C2
Platform	0.91	1	0.89	0.87	0.52
Time	0.95	0.97	0.88	0.87	0.67
Therapy	0.99	0.96	0.91	0.86	1
Subtype	0.92	0.94	0.90	0.81	0.80

Table 4.9: Silhouette Coefficients for Every Dataset Across All Integration Variables. Silhouette Coefficients measure two values, the difference from a sample to the mean of its cluster and the difference from a sample to the mean of other clusters. Combined these values are used to calculate the Silhouette Coefficient which describes the tightness and distinctness of a cluster with values ranging from -1, no clustering, to 1, strong presence of clustering. The values reported for these integration attempts all strongly signal that the clustering of the integrated data remains robust and the presence of batch is still felt in the expression data.

4.8 Results | Molecular Subtyping and Prognostic Scores

4.8.1 Pretreatment Samples with Platform ComBat Batch Correction

Subtyping and concordance of the resultant molecular/intrinsic subtypes is shown below in Figure 4.8.1, Page 112 for Pam50, rorS risk classification, MammaPrint risk classification, iC10 subtyping and ssp2003/6. Agreement of the calculated subtype is relatively high (concordance (subtype agreement) > 80%, Table 4.10, Page 110) for all tests except iC10 which was in agreement 72% of the time. This result shows that the concordance post integration is sensitive to the number of classifications possible for each test, where the greater the number of outcomes, the lower the concordance. These values show results in-line with patient-matched samples from different tissues or different treatment time points (Chapters 2 and 3), indicating that the distortion of *ComBat* is at least no greater than the effect of treatment. There are many colors on display here, to save wasteful repetition or detracting from the diagrams, I will outline the colors and matching typing for each test for the following diagrams. For the Pam50 and scmod1/2 classifiers there are LumA(dark blue), LumB (light blue), Normal (green), Her2 (pink) and Basal (red), for the SSP2003/6 tests they output ER-/HER- (red), HER2+ (pink) ER+/HER- High (blue) and ER+/HER- Low (green). The IC10 subtypes are depicted as: IC1 (orange), IC2 (slate blue), IC3 (dark seagreen), IC4 (deep pink), IC5 (azure), IC6 (dark orange), IC7 (khaki), IC8 (dodger blue), IC9 (goldenrod), IC10 (plum).

	Pretreatment only samples	Pretreatment Integration	Samples	Complete join	Platform	Time	Therapy	Subtype
C1	34	0.15	95	0.71	0.77	0.81	0.79	0.7
C2	141	0.23	248	0.79	0.85	0.84	0.79	0.83
E1	128	0.19	385	0.8	0.82	0.86	0.9	0.87
E2	58	0.22	176	0.54	0.625	0.76	0.57	0.73
E3	109	0.1	218	0.82	0.95	0.97	0.98	0.97
	Mean	0.18	**Mean**	0.74	0.8	0.85	0.8	0.82
	Average retention	82	**Average retention**	26	19.6	15	19.2	18

Table 4.10: Average Amount of Subtyping Agreement Loss Per Dataset for Every Integration Method. Each row contains the average loss of concordance for every data set for every test. The rows for mean represent each tests average performance, Average retention inverts the result to show the number of samples on average that experience no change. The first and third column shows the number of samples being compared in each test. The second, fourth, fifth, sixth and seventh show the proportion of samples that did change upon integration, and the bottom row reports the inverse average value which is the proportion that stayed the same through out.

4.8.2 Uncorrected Combination of Sequentially Sampled Datasets

For the data of the subset feature joins, no subtyping was available as the reduced gene lists were too incomplete to conduct these profiling tests. For the full outer joined data, comparisons were drawn for every sample from every dataset with the geneFu results from the unintegrated data, Figure 4.8.2, Page 113. Full comparison tables are available at `https://gitlab.com/rjbownes/Integration_Rationale/Compare_df`. A summary of the differences is presented below in Table 4.10, Page 110. As no assumptions were made as to the relative importance of the subtyping/classification tests, a non-concordant subtype in any of the geneFu tests resulted in a step change reduction in agreement. Agreement of the pre and post integration subtyping was low at a mean value of 26.8% unchanged samples. Additional rorS and MammaPrint continuous risk score comparisons were made, which clearly show a systematic decrease in the presented risk post integration.

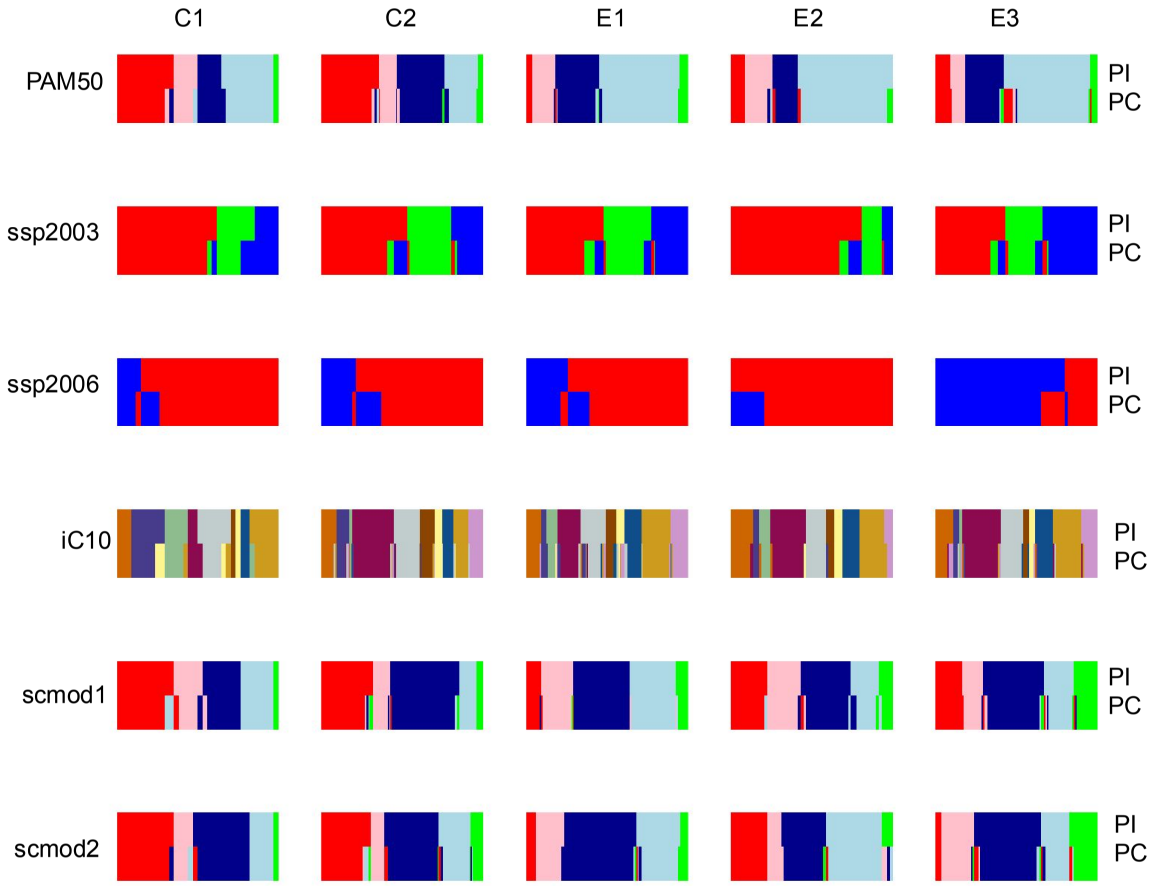


Figure 4.8.1: Subtype Concordance of Pretreatment Integrated Data Presented here are tiled plots representing the agreement between the pre-integration and post integration geneFu subtyping analysis results for six different tests; in descending order Pam50, ssp2003, ssp2006, iC10, scmod1, scmod2 and each column represents one dataset; C1, C2, E1, E2, E3. What is clear from an initial observation is that concordance between integration states is relatively high and is consistent between the data sets.

Figure 4.8.2: Subtype Concordance of Complete Feature Joined Data. Presented here are tiled plots representing the agreement between the pre-integration and post integration geneFu subtyping analysis results for six different tests on the pre- and post-treatment data; in descending order Pam50, ssp2003, ssp2006, iC10, scmod1, scmod2 and each column represents one dataset; C1, C2, E1, E2, E3. There is significantly more discordance in the subtype agreement than in pretreatment only integration.

4.8.3 *ComBat* Integration of Sequential Samples with Multiple Covariates

Analysis of the subtype composition changes for this integration based on the covariates resulted in a net decrease in concordance as compared to simple joins and to the integrated pretreatment data, Figure 4.8.3, Page 115 illustrates this point for the Pam50 subtyping test. The standout result is the large increase in the number of Luminal A samples post integration in the aromatase inhibitor treated cohorts and a small decline in the number of Basals in the Chemotherapy treated datasets. This is suggestive of a systematic shift in the centroid mapping algorithm of Luminal B samples relative expression of certain genes changing enough to be reclassified as Luminal A. Similarly and possibly more destructively, the Chemotherapy Basal samples appear to change most frequently to Her2-enriched samples. Table 4.10, Page 110 contains the number of undistorted samples per *ComBat* covariate method. *ComBat* performed worse than the Pretreatment only and uncorrected integration in terms of subtype concordance at 19.6%, 14.2%, 19.1% and 18% unchanged samples for Platform, Time, Therapy and Subtype integration respectively.

In addition to the categorical tests, continuous risk assessment of the samples were calculated post integration for each method, and compared to pre-integration values. Figure 4.8.4 highlights this result for the rorS prognostic test, on the integration with only the information of Batch in the pre and on-treatment data, concurrent results were obtained for each method, were highly concordant, and only one is being illustrated here for clarity. Full views of each comparison were performed but added little if anything to the analysis, and were only distracting as extra diagrams in text. As Figure 4.8.4 clearly shows via the loess regression lines comparing the pre and post integration prognostic values, there has been a significant shift in the relative risk scoring of the patients post integration (> 95% confidence intervals) for all datasets. E2 experienced a small drop in the perceived risk at lower rorS scores, but C1, C2, E1 and E3 all saw a statistically significant increase in risk for the better prognostic class patients ($p < 0.005$ for all datasets but E2 by pairwise Wilcoxon t-test). At high levels of estimated risk the five datasets show concurrent reductions in risk post integration. Regardless of direction this represents a large and systematic divergence of perceived patient prognostic characteristics.

Figure 4.8.3: Subtype Concordance of Pre- and On-Treatment Data Pam50 subtyping of the pre and on-treatment data before and after integration shows a significant difference between the concordance of the pretreatment only integration results (top row) and for each ComBat method (subsequent rows). This is especially obvious in E3 the large number of LumA now presenting as LumB. This does correlate with later results on the concordance of continuous risk scores.

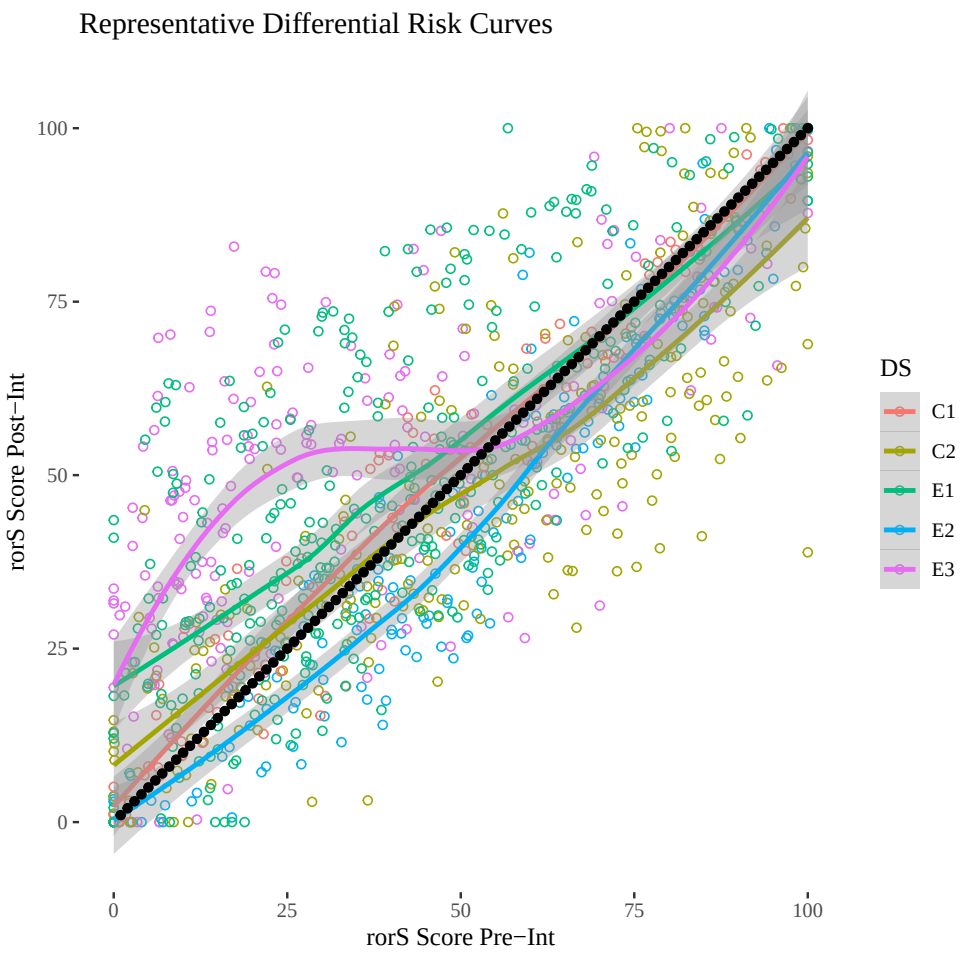


Figure 4.8.4: Continuous Risk Comparison Pre- and Post-Integration Comparison of the Pam50 rorS continuous risk assessment scores, here the x axis is pre-integration and the y is post-integration shows there is a large and systematic shift away from the pre-integration risk values. DS E3 shows a significant increase in the preceived risk of the low risk patients ($p = 0.0011$ paired Wilcoxon t-test for samples in Low or Intermediate rorS classifications pre-integration), this is consistent with the results of the subtype concordance scoring.

4.9 Results | Differential Expression Analysis

The inherent value of this data lies in the information contained in the pre- and on-treatment gene expression fold changes. These have previously been shown to help separate response classes to chemotherapy[215] and identify markers of response in aromatase inhibitor treated cohorts.[137] Maintaining these changes is of paramount importance to the value of this work. DE gene lists were calculated using linear Bayesian models and filtered for significance (FDR < 0.05). These were compared pre and post integration to evaluate the overlap in retained differentially expressed genes. This comparison is not drawn for the pretreatment only integration as there are no pairwise differences across treatment time.

4.9.1 Uncorrected Combination of Sequentially Sampled Datasets

Table 4.11, Page 117 contains the overlap values for the shared differentially expressed genes present in the reference unintegrated data and both uncorrected integration methods. Dataset E2 has the highest agreement with all the correction methods with the independently analysed data, however the remaining data share comparatively little with their references, with as little as 3.4% of the same genes being differentially expressed. In addition the complete overlap has ~13% more conserved similarity when compared to the subset feature integration.

	Differentially expressed gene list overlap pre- to on-treatment						
	Complete Join	Subset Join	Platform	Time	Therapy	Subtype	Mean
C1	21.2%	13.4%	22.6%	22.6%	23.4%	22.4%	20.9%
C2	32.4%	29%	27%	27%	31.8%	27%	20.9%
E1	42.4%	38.4%	48.8%	48%	39.4%	47.4%	44.7%
E2	91.6%	41%	50.4%	48.2%	87.4%	49%	61.27%
E3	3.4%	4.4%	5%	4.6%	4%	5%	4.4%
Mean	38.2%	25.5%	30.7%	30%	37%	30%	

Table 4.11: Percentage of Differentially Expressed Genes Retained After Integration. Each row contains the percentage of differentially expressed (DE) genes retained for each method, for each data set. The summary statistics are listed on the bottom and right most column.

4.9.2 *ComBat* Integration of Sequential Samples with Multiple Covariates

Results, Table 4.11, Page 117, are similar in magnitude to the pretreatment attempts at integration but do not represent an improvement in the retention of biological features, according to the differentially expressed genes. There is less

average similarity across the five datasets than before and a new minimum similarity (E3/4%). This strongly suggests that the important patient-matched translational changes that are a product of neoadjuvant treatment are being further lost to *ComBat*. Dataset E2 has the highest average agreement regardless of method, further identifying the outlier nature of this dataset.

4.10 Results | Random Forest Classification

Machine learning methods for the classification of cancers[244] and the prediction of recurrence of breast cancer[245] is well established. In particular, random forest has been shown to be a robust algorithm for pattern detection.[245] In this study, a random forest was trained on the pre integration data in order to try and identify the original batches in each post integration method. By inverting the results of the random forest output, we can gain a measure of the level of integration based on the mixing after *ComBat*. Ideally, this results in the "failure" of the RF model to correctly identify the dataset of origin, indicating that the descriptive features of each sample have been sufficiently shifted together to appear as if from one distribution

4.10.1 Pretreatment Samples with Platform ComBat Batch Correction

Random forest classification was 17% accurate (F1 score) in the identification of the original batch in the post integration data. This is below the 28% majority classifier performance and a clear indication of the successful integration of expression data. Integration of pretreatment samples with QN and *ComBat* show that features descriptive of batch have been eliminated and the distributions are shifted together well enough to represent a single batch.

4.10.2 Uncorrected Combination of Sequentially Sampled Datasets

Naive integration methods showed poor levels of integration according to the inverse RF metric. The random forest model classified 100% of the Complete feature joined samples correctly with the batch of origin this indicated that the integration was not sufficient to obscure the original batch effects. The subset feature joined samples had visible overlap of the E1 and E2 samples and this is represented in the RF model performance at 82% accuracy. Both of these methods failed to shift the underlying expression values sufficiently to render the batch of origin indistinguishable.

4.10.3 *ComBat* Integration of Sequential Samples with Multiple Covariates

Random forest mislabeled 142 samples (981/1122, 87%) for the platform integrated pre and on-treatment samples, 250 (872/1122, 77%) for treatment time integrated, 400 (722/1122, 64.3%) for the treatment type covariate and 478 samples (644/1122, 57.4%) for subtype adjusted *ComBat* correction. Subtype integration had the lowest random forest accuracy, correctly identifying the lowest number of integrated samples but still failed to reach the 28% benchmark. All *ComBat* assisted integration methods, except the Platform covariate, out performed the uncorrected method, however they all fell short of the pretreatment only integration performance.

4.11 Results | Proliferation Changes

As proliferation is an important biological marker of response to treatment, ensuring that the expected reduction to proliferation is still present post integration is an important metric to validating the integration methods. A reduction in proliferation markers has long been a marker for response to therapy[97] and to see a large deviation away from this would indicate a systematic disturbance to the fold change values in the pre and post treatment samples. Figure 4.11.1, Page 120 shows that for each dataset there is a significant reduction in proliferation markers on treatment compared to pretreatment expression levels in the data prior to integration. For all methods, the post integration on-treatment fold change values are not statistically distinguishable from zero but also have positive means. This indicates that universally, the starting values for the proliferation markers profiled in this test (PCNA, MCM2, MKI67, AURKA, FOXM1, BUB1 and TOP2A) have been reduced with respect to the patient-matched on-treatment values or the post treatment samples appear to now have higher expression than before. The implication of this result however is that the changes on-treatment no longer represent the known, normal biological trend of the pre-integration data and are no longer representative of of these important markers. These values are tabulated in Table 4.12 below.

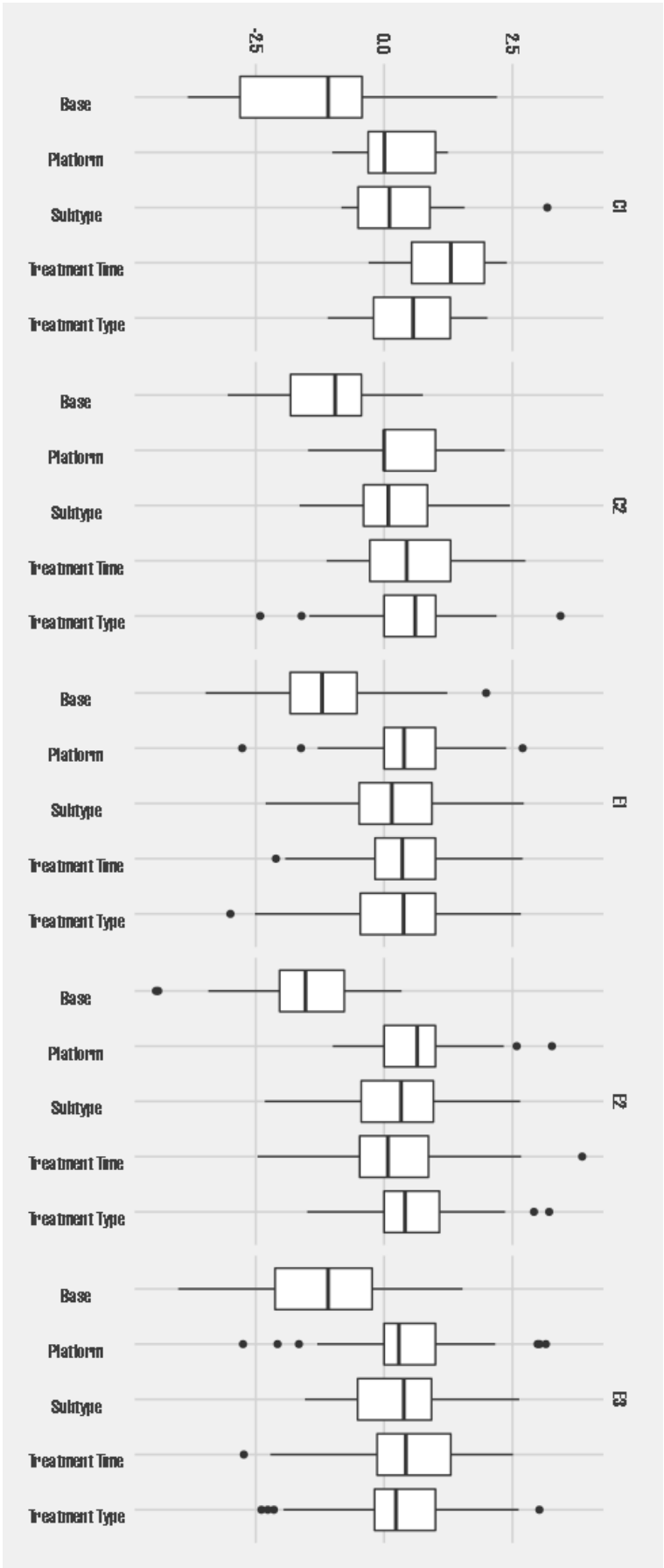


Figure 4.11.1: Changes to Normal Proliferation Values On-Treatment. Proliferation is normally reduced on-treatment. This figure shows that the normal reductions in proliferation are distorted post integration for all methods, here the "Base" represents the normal reductions in proliferation in the uncorrected, independent, datasets and each other feature illustrates the change in proliferation for the ComBat correction methods of the pre- and on-treatment data.

	Proliferation Change Pre-Int	Platform Int.	Subtype Int.	Treatment Time Int.	Treatment Type Int.
E1	-0.83(-1.90, -0.22)	0.42(0.02, 0.61)	0.18(-0.24, 0.57)	0.37(-.12, 0.58)	0.38(-0.21, 0.60)
Sig. Diff. from Zero.	0.0069	0.98	0.91	0.90	0.92
E2	-0.84(-1.91, -0.33)	0.46(0.0, 0.61)	0.23(-0.24, 0.55)	0.10(-0.25, 0.51)	0.20(0.01, 0.49)
Sig. Diff. from Zero.	0.0019	0.90	0.93	0.91	0.89
E3	-0.71(-2.66, -0.21)	0.22(0.0, 0.47)	0.26(-0.33, 0.42)	0.28(-0.10, 0.71)	0.19(-0.12, 0.44)
Sig. Diff. from Zero.	0.0081	0.93	0.92	0.92	0.91
C1	-0.70(-2.51, -0.18)	0.03(-0.11,0.82)	0.09(-0.19, 0.73)	0.97(0.52, 2.11)	0.51(-.16, 0.91)
Sig. Diff. from Zero.	0.0089	0.92	0.91	0.92	0.91
C2	-0.78(-1.88, -0.22)	0.01(0.01,0.89)	0.11(-0.19, 0.79)	0.42(0.22, 0.91)	0.49(01, 0.65)
Sig. Diff. from Zero.	0.0078	0.94	0.92	0.96	0.90

Table 4.12: Table of Values Representing the Changes to Normal Proliferation Values On-Treatment. The proliferation fold change values for the aggregated genes shown in the above diagram are represented here. Each column represents one on of the integration methods, the mean and interquartile range is presented as well as the standard T-Test result for comparing the distribution of this data from a fold change of zero.

4.12 Results | Post-ComBat Distributions

To further investigate the possibility of the on-treatment samples causing the additional noise, Figure 4.12.1, Page 122 was drawn to highlight the *ComBat* process of distribution normalisation. This diagram shows the contributions of each additional dataset to the complexity of the normalisation calculations. Each dataset introduces new distributions that must be shifted and in turn these increase the complexity of the integration function. It is possible to see that there are additional "shoulders" in the distributions of the pre and on-treatment data, indicating bi or multimodal distributions in the data, which are partially obscured by larger or wider distributions.

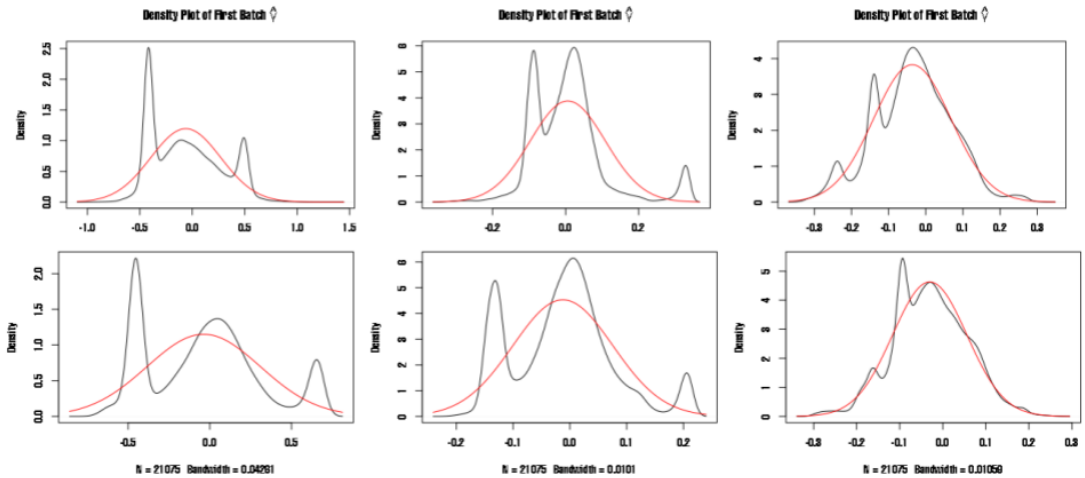


Figure 4.12.1: Evidence of Multiple Incompatible Distributions in the Pre- and On-Treatment Data. It is possible to discern the shoulders of additional subdistributions in each dataset for the different combinations of datasets that are likely antinomic to the ComBat distribution shifting process. This diagram is meant to visualise the distributions that are being normalised and combined togther in the ComBat process. As can be seen from the successive addition of each new batch, the columns, there is increased topological complexity to the underlying data. The order of addition of the data sets changes the mid-step topology, but the final conformation is always the same. In this diagram the top row represents the non-parametric, and the bottom the parametric, implementations of ComBat. The first panel shows three datasets, E1/2/3, the middle panel adds C1 and the final panel adds C2.

The presence of additional sub-distributions was further explored by examining the distributions of the gene expression on a per-time point-per-dataset level.

Figure 4.2.2, Page 76 shows the mean centered gene expression distributions for every dataset (A), then the sub distributions of treatment-by-time point as box plots (B), and kernel densities (C). There are clear distinctions between the time-point-treatment distributions, as seen by the box plots and kernel densities, both of which support the idea that these additional distributions are limiting the ability of *ComBat* to successfully integrate the data.

4.13 Discussion

There were an enormous number of possible covariates of *ComBat* integration; dataset, sequencing platform, type of treatment, time on treatment, intrinsic subtype, known response, histological grade, ER positivity, PR positivity and HER2 status, detection method, BRCA mutation status, node status, age, tumour size and menopausal status to name just a few. Of these, the factors that were trialled were platform, treatment time, treatment type, and intrinsic subtype. These features were chosen as covariates to improve the performance of *ComBat* primarily as they were the most feature complete variables for consideration and would retain the greatest sample number. Ideally, these factors would account for most of the inter-dataset variation and enable the most uncompromising integration of the data.

Integration of these disparate cohorts has proved an enormously difficult undertaking. Batch correction is a fairly standard procedure,[246] but is possibly insufficiently subtle to combine on-treatment neoadjuvant data and keep the important and highly sensitive patient-matched fold change values intact, despite its use in the normal removal of these effects from microarray data.[247,248] Kernel density estimates of the normalised data for each integration method suggests that integration can be performed completely as the distributions of the expression data appear as a singular, uniform, curve with normal parametric characteristics. Importantly, each component distribution is fully indistinguishable from the others. This is strongly juxtaposed by the results of the genewise clustering, principal component analysis, subtyping concordance, random forest analysis, correlation testing and analysis of the shared differentially expressed genes which all would indicate that the integration of the data has been highly descriptive to the underlying data and thus totally unsuccessful. *ComBat* has previously been shown to improve the statistical analysis of -omic data and improve the identification of differentially expressed genes,[249] and concordance with known biological pathways[250] but that seems to falter with the addition of the on-treatment samples.

Integration of single time point cross platform data, has been previously described[216,241] and successfully performed with quantile normalisation and *ComBat*. As expected, the post integration results of this method were the most concordant. Correlations of the patient samples post integration were comparatively high, > 75% and PCA results of this protocol were visually conforming to the expected integration outcome. These two tests indicate that the patients still resemble their pre-integration expression profiles, but the datasets are no longer distinct. Subtyping and risk scoring showed similar levels of overall similarity with ~80% subtype agreement post integration and a non-significant difference in the presented risk of recurrence. This is especially important as these metrics are valuable descriptive features with implicit biological and clinical implications that need to be maintained to ensure fair analysis post integration. Discrepancies here would indicate that the post integration data does not have the same biological representation as the pre-integration data. Because the resultant subtyping and risk assessment tests are so concordant, inferences made on the post integration data can still justifiably applied to the pre-integration samples allowing for integrative analysis. Lastly, supervised machine learning models were unable to distinguish the study of origin in the pretreatment only integrated samples. Overall, this signifies that the biological diversity of the original data was meaningfully maintained, while simultaneously removing as much of the technical variance as possible, this is certainly in keeping with the results from Turnbull et al. [216] and other modern examinations of single time point integration retaining significant biological information.[249,250]

Uncorrected approaches to integration were outperformed by the established pretreatment only protocols. Correlations of the cross integration matched patients samples were lower as a whole with the exception of dataset E2, which was aberrantly well correlated. Post integration the overall expression values were not as representative of the pre-integration data, suggesting more disturbance to the relative gene expression values possibly as a result of the lack of subtlety in the basic joining methods. The subsequent PCA diagrams illustrate that, regardless of the change to the underlying relative gene expression, the inter-dataset variance is maintained. This result indicates that no technical variance has been removed but the biological values have been altered. Subtype agreement post integration was very low, just 26.8%. This confirms the result of the correlation analysis that the relative gene expression values have been destructively shifted. The post integration subtyping and risk assessment values are now no longer representative of the un-integrated data making subsequent

analysis of the data irresponsible from this perspective. Additionally, the pairwise differentially expressed genes were comprised of almost entirely different genes (38.2% and 25.5% agreement), meaning the valuable treatment-dependent fold change values have also been compromised. Lastly, the ML model was successful at classifying the dataset of origin, thus confirming the results of the PCA analysis where the inter-dataset variance and differentiating features have been maintained but the important biological characteristics have been lost. In context, these results are rational and expected as corrections have long been established as appropriate protocols for improving cross platform analysis.[251]

ComBat integration with multiple covariates was outperformed by the pretreatment only integration and was roughly on par with the uncorrected integration methods. Correlation analysis revealed further reduced patient matched similarity ~60%, again with the exception of dataset E2. This suggests there is a systematic failure of these methods to alter the general expression profiles of this dataset, however subsequent analysis fails to reveal a sufficiently viable reason. Heatmap and PCA analysis of these methods complement each other to emphasise the fact that the datasets are not significantly integrated on any of the highlighted clinical or technical factors. Additionally, any overlapping of the samples in the heatmaps can be seen more clearly in the PCA as non-integrative dispersion of the data, which scales with the number of covariates. This is a clear illustration of the fact that *ComBat* assumes uniform contributions of the covariates, a feature that works against very heterogeneous data. The inability of *ComBat* to sufficiently handle these non-uniform distributions is shown in the analysis of sub-distributions in these cohorts, which demonstrate clear and present multi-modality. Subtyping concordance was on par with the uncorrected methods, even when starting subtype composition was taken into account, which strongly indicates that the addition of covariates to integration is adding little value to this analysis. The pairwise fold change values were maintained only ~32% of the time, meaning the intrinsic values of these samples is again lost with these methods. Dataset E2 continues to be an outlier in this analysis, with an average retention of 61.27% of the DE gene lists. Lastly, the unsupervised machine learning methods had a range of performance in dataset classification. Fundamentally, this result would indicate an improvement in integration performance from `Platform < Treatment time < Treatment type < Subtype`, however with regards to the PCA and heatmap analysis this could equally be explained by the non-integrative dispersion and possible model over fitting. There are very few examples of analysis like this and correspondingly it is very difficult to place the results outside of this thesis. Muller et al. have previously shown that longitudinal integration is possible using precisely the methodolo-

gies employed here, but the successful integration of this data is not replicated in this study.[252] This is possibly due to the nature of the data being substantially different, this data is comprised of multiple time points from five different platform's, the Muller study was of a patient cohort with matched patient samples after five years of follow up on the same platform.

GeneFu was heavily utilised for this section and its analysis. This library provides easy to employ, reliable, means of classifying subtypes and predicting prognostic risk. However, geneFu and its methods are only approximations of the tests it seeks to emulate. They are in actuality models that estimate the classifications that would be conferred on to BC samples tested with laboratory conducted gene panels and IHC tests. This incurs a loss of accuracy inherently as they are derived from expression matrices and not the true tissue samples. geneFu primarily operates as a centroid mapping algorithm and this confers a second approximation that must be considered. Fundamentally this method works by comparing the transformed gene expression scores of pre-defined lists of genes on a per sample basis to point values mapped to subtype annotated centroids. Each sample is then compared to the centroid values and the classification of least distance is chosen as the output. This means that a sample can be massively dissimilar to every class, but will be assigned to the least distant one, and also these values are affected by dataset composition through normalisation and pre-processing, meaning the individual sample is affected by the composition of the whole cohort. These factors must all be considered fully when using this library, and that is why these values were primarily used for comparative analysis of the changed that occurred to integration instead of as an examination of the composition of each cohort.

Of special interest with dataset E2 is that the samples from this cohort always only cluster/align with samples from E2 despite also containing biological overlap with samples in E1. The same patient samples analysed under different conditions still do not appear more similar after integration. This suggests that the effects of batch are possibly insurmountable. This result stems from an error in the construction of the backend data structure but is a serendipitous result as it adds additional weight to the result that integration is not sufficient to over come the problems of batch.

4.14 Conclusion

Integration of sequentially matched neoadjuvantly treated datasets is a formidable obstacle to large scale on-treatment breast cancer analysis. Under the existing standards of data integration, post combination data was not representative of the pre-integration data. This results strongly suggests that a meta analysis of neoadjuvant therapies is a more appropriate method of comparing different datasets in a non-destructive manner. At this time, the standard methods for batch correction and dataset integration via quantile normalisation and *ComBat* correction are inappropriate for patient-matched, neoadjuvant data.

5 | Meta-Analysis of Multiple On-treatment Datasets Reveals Common Transcriptional Differences In Non-responsive Tumours

5.1 Abstract

Background

Analysis of small cohorts of neoadjuvantly treated breast cancer has yielded some initially powerful results in the area of biomarker identification and risk stratification of patients. However, these results are undercut by the limited sample size and possible lack of population representation of the underlying data. It may be possible to identify more subtle signals in integrated data and make inferences with significant statistical backing using the meta analysis of multiple studies of this scarce data.

Methods

Five datasets of transcriptomic breast cancer data and matched annotation information were collected and combined to create a repository of independently but homogeneously processed data. The crucial pairwise differences in gene expression were harvested to identify uniform patterns of change in gene expression common to all treatment types. Machine learning (ML) models including Support vector Machines and Neural Networks were trained to predict response from the pairwise fold change values.

Results

Analysis of the aggregate fold change values reveals that there is definite clustering of the patients based on the factor of non-response regardless of the treatment involved. Subsequent pathway analysis of these non-responsive samples reveals a preponderance of genes associated with, and enrichment of, mTOR and PI3K pathways, suggesting a possible commonality in nonresponse and new targets and treatments for these patients. ML models were trained to attempt to identify response status of BC patients cross treatment with significant accuracy (89% Support Vector Machine (SVM), 92% Neural Network (NN)).

Conclusion

Examination of matched pre and on-treatment samples facilitates new kinds of meta analysis from neoadjuvant data, yielding previously unseen insights into BC treatment response vectors. While endocrine therapy and chemotherapy have significantly different mechanisms of treatment and this can be seen in the way patients respond to treatment. The vectors of non-response appear to have overlap and are enriched for genes currently being targeted for new drug development, mTOR and PI3K. Lastly, the ML methods presented are able to leverage these values to train models for the accurate prediction of response across treatment paradigms by targeting these vectors of non-response.

Overview

A generalized diagram of for this work, especially with regards to the generation of the most vital outcomes is presented in Fig. 5.1.1, on page 131. This diagram highlights the critical difference presented in this chapter compared with the study of the expression level integrated analysis performed in the previous results section. Results from the individual analysis of different data sets is presents and the meta-analysis of the aggregated output is what is presented in this chapter.

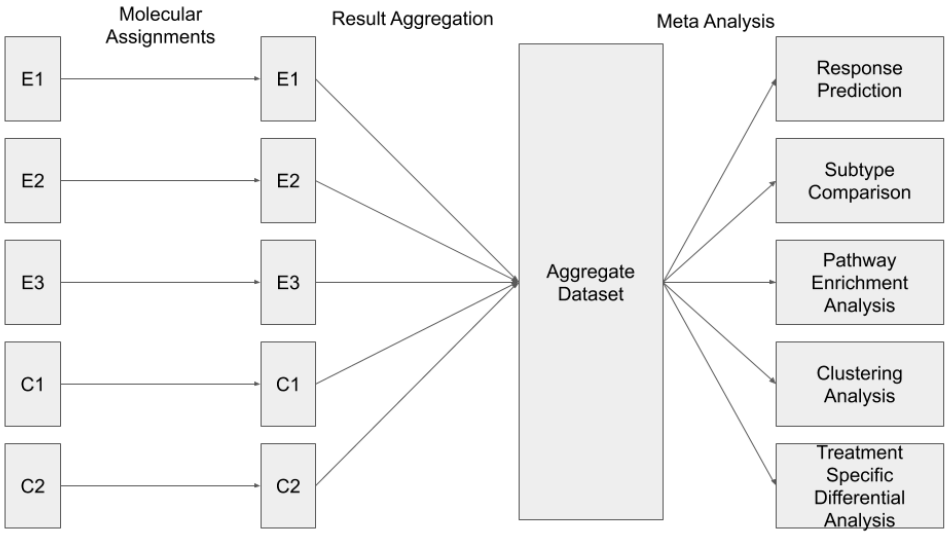


Figure 5.1.1: Overall Study Design and work flow. The workflow for this chapter is presented in this diagram. The datasets are seperat ely analysed for subtype, prognostic score and other molecular typings before the output of these results are integated to form a meta-dataset of viable and comparable results. Meta-analysis of these outputs is then then conducted to attain pan-treatment, cross dataset results.

5.2 Background

There have been previous attempts to to use integrated expression data *en masse* to enhance the significance of research and provide more powerful answers to questions posed about patient response prediction and risk stratification.[253,254] Multiple algorithms have been developed for correcting the "batch" effect associated with disparately sourced data and aggregating it to facilitate new research, these include Surrogate Variable Analysis (SVA),[218] Bayesian Factor Regression Modelling (BFRM),[219] Factor Analysis (FA)[220] and Array Generation Centering (AGC).[221] These methods all seek to shift the distributions of expression towards one another and by so doing, integrate and normalize the data. These methods are all different implementations which seek to evaluate the contribution of different features or variables that contribute to the shape of each different distribution in the parent data. Then, upon integration center the data as much as possible between the different data sources while maintaining the shape of the distribution of the identified important features. *ComBat*[222] has been established as the standard method for batch correction and integration of microarray data,[223–225] but is, as of yet, untested in matched patient sequential samples. However, due to the overwhelming evidence to suggest it's applicability here, it will be pursued over the alternatives.

Recent successful attempts at leveraging the results of multiple translational studies together have utilized various methods to escape these problems. Jiang *et al.* constructed a network based approach that individually assessed upstream involvement of gene expression to identify driver genes to describe specific gene expression profiles.[255] This is a popular approach, as there are several example network analyses of transcriptomic data with associated genomic quantitative and annotation data by other researchers from the last few years. Primarily, these studies utilize microarray data from multiple sources, but side step the problems of batch.[256] Other teams have focused on the integration of expression data with other types of data to perform intricate multi-omic analysis.[257,258] Tools to streamline the collection, creation, and annotation of databases from publicly available repositories have even been produced, however the tight integration of expression level data is uncommonly attempted and more rarely successful.[259]

To circumvent the problems of batch and multiple expression set alignments, I have created a unique set of comparable matched pre and on-treatment expression level change values to make summary analysis of the effects of neoadjuvant therapy. Here, I examine the results of the scaled pairwise differences in gene expression to gain new insight into the responses and changes in breast cancer to treatment.

5.3 Methods and Materials

The five datasets used in this analysis were gathered either from the public domain, or through locally hosted and maintained sources, details on each dataset can be found in Table 4.1 for annotation and summary statistics of samples including platform and treatment type. These were independently collected and pre-processed to remove probes with no signal or variance and quality control the samples with Limma,[230] DESeq[260] and edgeR.[231] The datasets were voom normalized independently and attempts were made to integrate the expression data using *ComBat* as described in Chapter 4. As these attempts were unsuccessful, MetaMatchBreast (MMB) was constructed using R[145] and the Bioconductor package XDE[236] establishing an ExpressionSetList object type to build a frame work of un-merged expression data with integrated annotation information. This provided an iterable list object for analysis and testing, with unified pathological and clinical information for sub-setting of the data.

Additional factors included are all available clinical features and follow up,

up as well as calculated subtypes. In addition, functionality for calculating dataset specific differential expressed based on defined subclasses from within the annotation has been written to allow for bespoke analysis of the bundled data. Integration methodologies are also built in to allow for the combination of any subset of the data with multiple different covariates to independent analysis. MMB also contains high level exploratory and visualization functions for survival analysis, statistical testing, comparative expression analysis and appropriate plot out-puts for these functionalities. Critically for the longevity of the work presented in this thesis MMB is infinitely extensible. The structure of MMB is such that the expression set list can be appended with any new expression matrix and the matching annotation information can be combined with the expression-set pheno-data object. This functions as a means for continuing to incorporate and grow this aggregate resource as new data becomes available.

The ability to extend the dataset to incorporate new and emerging neoadjuvant trial data is pivotal to the long term value of this work. As has been stated previously in the results chapters, this type of data is more rare and comparatively smaller due to the complications of multiple sampling. The ability to aggregate this data for meta-analysis will hopefully allow for enhanced future analysis of these samples to make more powered and profound results possible. The full scope of this work was not completed by the conclusion of this thesis work but has been carried on by myself and a new PhD student from my supervisor. Ideally, this will ensure the continued legacy of the work and provide stable maintained and updates for this work farther in to the future.

Fold change values of the pre and on-treatment samples' gene expression values were used to perform the majority of the analysis as the original data could not be directly compared. Differential analysis of the fold change values between treatment times of patients with known response (PCR vs. Non-PCR) was performed for candidate gene analysis using Limma to perform the DE analysis with a standard fdr of 0.5, this method uses Bayesian inferences about the probability of the gene association to response class. Additional ranked product analysis was performed using RankProd[261] in R. Clustering of the patient-matched fold change values was performed with T-distributed Stochastic Neighbor Embedding (t-sne) to try to identify response associated expression profile changes. Differentially expressed gene lists were then used independently of each other, and secondly using the overlap of the gene lists, for GSEA[262] analysis and CAMERA[263] pathway testing to better understand pan-treatment responses to therapy. Over-represented genes in the GSEA and CAMERA analysis were col-

lected and examined for correlated KEGG[264] pathways. This was done by plotting the representation of genes in their respective KEGG pathways in ascending order to identify an inflection point, an elbow, in the graph which signals the genes of lesser pathway prominence to the genes of greater biological relevance. The genes and pathways past this inflection point are selected for and brought forward for additional analysis. Tests of significance between expression level for groups or genes defined by this analysis were performed with standard Wilcoxon t-tests using a p value cutoff of 0.05 to indicate significance.

Python[265] was used for the ML model building which was trained using sklearn[266] as the common api to access the SVM[267] model, parameters, and for hyperparameterization. Tensorflow[268] and Keras[269] were used for building, compilation, and training of a sequential convolution neural network for the prediction of response class from the pairwise fold change values. The SVM model was trained with a radial kernel in order to better capture none linear boundaries between response classes. The neural network was trained with two dense layers with 64 neurons and a ten percent drop out layer to prevent over fitting. RELU activation functions were used at each layer with a softmax output and using a stochastic gradient descent loss function. The neural network and SVM model were trained on fold change values for all samples with known response in an 80:20 split of the data with a 10 fold cross validation of randomly sampled data. Subsequent analysis was performed by randomly removing one of the datasets, training on the remaining four and validating in the fifth in order to test the impact of the model accuracy on a per dataset level. Lastly, a pre-trained ResNet model[270] was used inside a pytorch[271] training loop to classify heatmap images of the before and after treatment samples.

5.4 Results

5.4.1 General Changes in Gene Expression are Conserved Between Treatments

Differentially expressed genes were identified as previously described[215] and in the methods with an fdr of 0.05 from the matched pre and on-treatment samples in the Chemotherapy cohorts. The genes lists defined by this method were probed in the endocrine treated cohorts to view the similarity in expression level change. This analysis shows similar patterns of expression in both cohorts, however with statistically significant differences (means expression level values of the average gene expression per treatment cohort $p < 0.05$ Wilcoxon t-Test) in

the amplitude of change, this indicated that the direction, but not magnitude of change across the treatment types was the same. This means that genes that become down-regulated post treatment in the Chemotherapy patients also see a reduction in expression in the endocrine treated patients but to a greater extent. This indicates that the response to treatment is similar but possible pathway involvement and specific methods of action are different between the two treatment types. This result shows that many of the same changes occur in Chemotherapy as in Endocrine therapy and separation of response or treatment specific changes will need to be examined more subtly.

Following this analysis, rank product analysis between the response groups of each treatment type was performed and the results can be found in Table 5.1. This found 46 genes that were differentially expressed between the response groups that overlapped in both treatment types. Subsequent analysis however found no disease specific pathway enrichment from this gene list, indicating that this is probably a statistical anomaly presenting false confidence in this overlap as there is no apparent biological drivers to this shared differential expression.

RP analysis of Chemotherapy Samples					RP analysis of Endocrine TherapySamples				
Gene	RP/Rsum	FC:(class1/class2)	pfp	P.value	Gene	RP/Rsum	FC:(class1/class2)	pfp	P.value
ATP6V0C	1	2.52	9.70E-07	5.17E-10	DEFA3	1.21	2.6	9.64E-06	9.68E-09
KLF9	1.24	2.21	8.93E-07	4.88E-10	ADCY3	2.53	2.39	9.34E-06	8.71E-09
CDH11	3.04	2.2	8.92E-07	4.13E-10	LAMP2	2.81	2.38	6.78E-06	7.05E-09
FANCB	3.63	2.19	8.80E-07	3.33E-10	CCNG2	3.97	2.34	9.06E-05	4.23E-09
MBP	4.75	2.18	4.26E-07	3.31E-10	FANCB	4.56	2.31	8.27E-05	8.12E-08
EYA4	5.64	2.14	9.52E-06	8.00E-09	AQP3	4.62	2.29	7.31E-05	7.76E-08
F2	6.1	2.13	8.45E-06	7.28E-09	F2	5.68	2.27	6.61E-05	6.92E-08
TCF19	7.14	2.1	8.05E-06	6.98E-09	USP1	5.69	2.26	6.60E-05	6.76E-08
DEFA3	7.58	2	7.00E-06	6.93E-09	FOXM1	5.95	2.24	5.99E-05	6.20E-08
FOXC2	8.62	1.99	3.34E-06	6.52E-09	CHD2	7.84	2.05	5.62E-05	5.54E-08
HGF	8.9	1.88	2.18E-06	5.55E-09	GSTM3	8.17	2.01	8.86E-04	4.99E-08
JAG2	10.6	1.87	9.35E-05	9.75E-08	HLA-C	8.42	1.99	8.76E-04	7.78E-07
LGALS3BP	11.11	1.75	7.49E-05	9.27E-08	ITGB3	8.58	1.9	8.21E-04	7.76E-07
LIG3	11.13	1.63	6.51E-05	8.70E-08	RFC1	9.58	1.85	7.25E-04	7.00E-07
EPAS1	11.8	1.56	6.33E-05	8.68E-08	MAP4	11.33	1.82	6.64E-04	6.60E-07
MYO10	13.52	1.55	5.96E-05	8.12E-08	NR3C2	11.54	1.79	6.28E-04	5.88E-07
NGFR	13.8	1.52	5.12E-05	7.35E-08	MYO6	12.98	1.78	6.12E-04	5.42E-07
SERPINF1	14	1.5	4.51E-05	7.24E-08	ACTB	13.48	1.76	2.97E-04	5.31E-07
PRKAG1	14.01	1.48	4.34E-05	6.36E-08	SERPINB8	13.57	1.74	7.76E-03	4.28E-07
CCL18	14.29	1.46	9.35E-04	8.29E-07	NUMA1	14.66	1.71	6.96E-03	9.62E-06
PURB	14.34	1.43	8.40E-04	7.50E-07	PHF1	14.81	1.69	6.05E-03	8.86E-06

Table 5.1 continued from previous page

RP analysis of Chemotherapy Samples					RP analysis of Endocrine TherapySamples				
RARG	14.96	1.41	7.98E-04	6.51E-07	NDUFS4	15.08	1.68	5.52E-03	8.73E-06
UQCRFS1	15.09	1.4	7.24E-04	5.98E-07	GATB	15.7	1.65	4.94E-03	8.45E-06
RPS6KA2	15.28	1.39	6.50E-04	5.26E-07	PROS1	15.76	1.62	4.28E-03	7.00E-06
RREB1	15.78	1.36	5.68E-04	2.18E-07	KLK7	15.94	1.62	3.31E-03	6.21E-06
PSMD4	15.98	1.3	9.98E-03	9.84E-06	GAA	16.03	1.59	2.88E-03	5.28E-06
SELE	16.53	1.29	9.84E-03	7.34E-06	RPS3A	16.28	1.56	2.54E-03	4.50E-06
SNTA1	16.66	1.28	8.56E-03	4.05E-06	RREB1	16.48	1.54	8.20E-02	4.06E-06
FABp5	16.79	1.25	8.13E-03	3.49E-06	SLC5A3	16.73	1.53	7.53E-02	8.53E-05
TEF	17.15	1.17	7.82E-03	2.94E-06	NGFR	16.9	1.52	6.76E-02	7.57E-05
CCNB2	17.72	1.15	6.88E-03	2.34E-06	SQLE	17.1	1.49	5.98E-02	6.96E-05
TPD52	17.99	1.06	6.24E-03	1.62E-06	STAT1	17.56	1.44	5.58E-02	6.94E-05
TRPM2	18.92	0.98	5.28E-03	1.52E-06	UGP2	17.93	1.43	4.23E-02	5.98E-05
UGP2	20.65	0.97	4.29E-03	1.40E-06	SPR	18.1	1.4	3.42E-02	5.42E-05
RBP1	20.84	0.96	3.36E-03	1.38E-06	VEGFC	18.35	1.39	2.71E-02	5.25E-05
ZNF24	20.89	0.94	3.20E-03	1.32E-06	ZAP70	18.57	1.37	2.50E-02	4.52E-05
SRPX	20.95	0.89	3.15E-03	1.21E-06	REEP5	19.04	1.33	1.92E-02	3.75E-05
RIPK1	21.24	0.85	2.14E-03	1.19E-06	ALX1	19.42	1.27	1.45E-02	3.58E-05
ASAP2	21.89	0.8	2.13E-03	1.11E-06	SRPX	19.57	1.24	1.38E-02	2.99E-05
PER3	22.62	0.77	9.91E-02	9.08E-05	NDST2	19.81	1.23	1.32E-02	2.87E-05
PSTPIP1	22.9	0.73	9.49E-02	3.35E-05	PTCH2	20.32	1.19	1.28E-02	2.53E-05
TRAPPC10	23.95	0.72	8.62E-02	2.49E-05	STK19	20.53	1.13	1.25E-02	1.58E-05

Table 5.1 continued from previous page

RP analysis of Chemotherapy Samples					RP analysis of Endocrine TherapySamples				
VAMP3	24.02	0.7	7.75E-02	1.24E-05	EIF2B3	20.6	1.1	1.17E-02	1.31E-05
MINPP1	24.03	0.6	7.08E-02	1.16E-05	RNF8	21.75	1.01	1.13E-02	1.24E-05
DHRS9	24.21	0.57	6.31E-02	1.11E-05	ZMYM6	22.39	0.88	1.12E-02	1.11E-05
PEMT	24.9	0.55	5.27E-02	1.08E-05	MED27	22.42	0.85	1.07E-02	1.07E-05
					LPIN2	22.83	0.8	1.06E-02	1.05E-05
					MAML1	23.31	0.78	1.05E-02	1.02E-05
					IP6K1	23.53	0.77	8.62E-01	8.64E-04
					GINS1	23.93	0.76	7.00E-01	8.32E-04
					ABCA10	24.02	0.69	5.54E-01	6.62E-04
					PEMT	24.17	0.65	4.99E-01	4.64E-04
					ATG7	24.27	0.54	3.00E-01	4.27E-04

Table 5.1: Table of The Rank Product Analysis of The Non-responsive and Responsive Samples in Both Treatment Cohorts. This table contains the Entrez gene ids, as well as the ranked product/ranked sum value, median fold change value, sample pFP score and p-value. The pFP value is the estimated percentage of false positive predictions when both up and down regulated genes are considered. This shows the data for the 46 differentially expressed genes between the non-responsive and responsive samples in the chemotherapy cohort and the 53 genes identified in the endocrine treated samples.

5.4.2 Non-Response Vectors of Chemotherapy and Endocrine Show Concordance On-treatment

Pre to on-treatment fold change values of matched patients cluster separately on the basis of treatment and response when viewed as feature reduced t-distributed Stochastic Neighbor Embedding (t-SNE) representations. t-SNE is gaining popularity as a dimensionality reduction and visualization technique for human genomics studies, it is potentially more sensitive for very high dimensionality data, like is presented here, with regards to subpopulation stratification than PCA.[272] PCA and t-SNE are comparable methods in that they are both methods of dimensionality reduction and make visualization of very high dimensional space possible in two dimensions for regular plotting as in the below example. They differ in that PCA is a deterministic model that measures the variance of orthogonally derived vectors which are composed of combinations of the data's features and t-SNE is a stochastic (probability based) model that measures the likelihood of samples being close together in space. Responsive Chemotherapy and responsive Endocrine therapy patients clearly cluster away from the non-responsive samples, which appear to have a more similar lateral and vertical displacement, Figure 5.4.1, page 140. This is possibly suggestive of differing pathways of response to different therapies but common pathways of non-response. This possibly indicates that non-responsiveness can be predicted using the pairwise differential gene scores.

To enumerate the differences between the response classes, I plotted the D1/D2 contributions (these are the vectors that describe the lowest space reductions, dimensions (D) that the data can be fit to, Dimension 1 and 2, or when plotted commonly X and Y) for each sample and made statistical tests (Wilcoxon T-test) between each group to see which were significantly distant, Figure 5.4.2, page 141. In the first dimension, the non-responsive samples are indistinguishable from each other but are significantly separated from the responsive samples from either treatment paradigm. In the second dimension, there is significant separation between all response-treatment classes. The statistical differences presented in this analysis are highly suggestive that there are systematic differences between the non-responders and the other classes and between the response classes. I was unable to repeat this analysis for different clinical factors as annotation is not uniformly complete for relevant clinico-pathological factors (grade, receptor status, etc.) and would have been impossible to compare the significance of results between different factors and sample sizes without compromise.

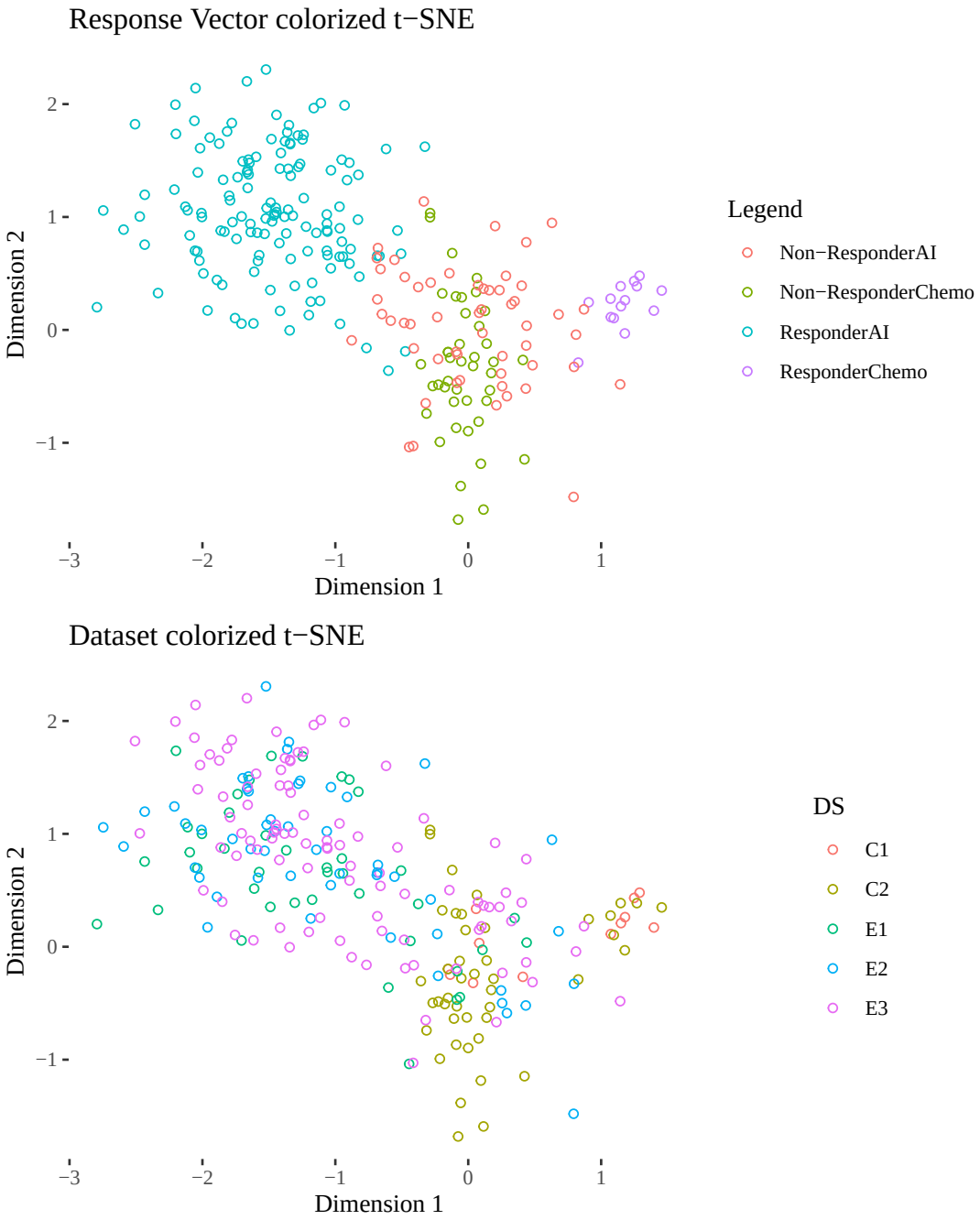


Figure 5.4.1: t-SNE of BC Transcriptional Pairwise Values Shows Clustering of Patients by Categorical Response. t-SNE shows clear clustering of groups by the joint grouping of response and treatment but with overlap of the non-responsive endocrine and chemotherapy patients. t-SNE is a probabalistic method of expressing similarity, hence this results suggests that the groupings visible here are due to inherent similarity of the samples in each cluster. The data in this model are the pairwise expression level changes of the on-treatment samples relative to the pre-treatment.

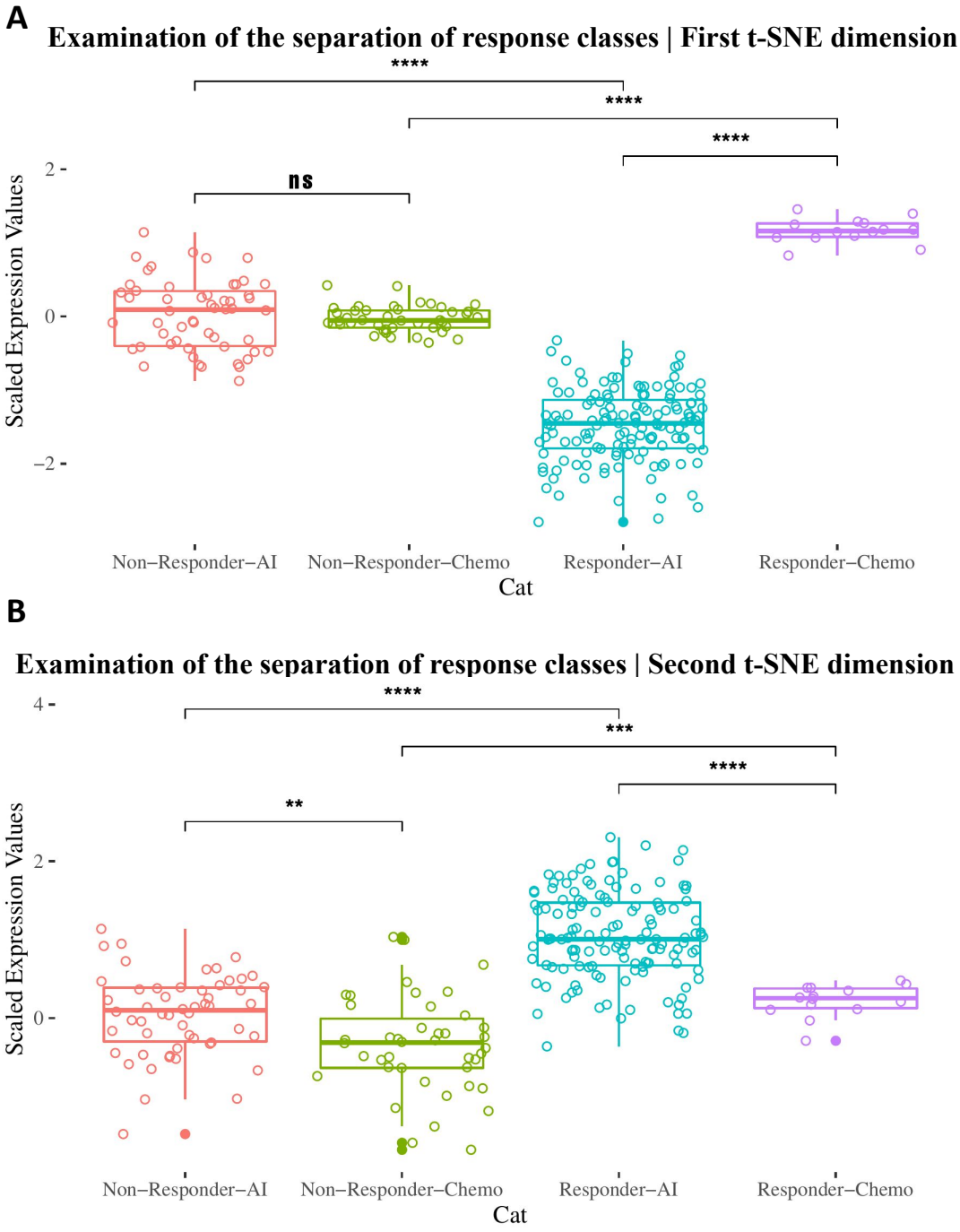


Figure 5.4.2: Statistical Analysis of Intergroup Differences Show Significant Seperaration on Both Principal Components. Significance scores (Wilxocon T-test) for the intergroup means differences are presented with stars. AI is used to denote the endocrine treated cohorts for brevity in the visualization. A) shows the intergroup differences in mean expression of the eigenvetors that comprise the first dimension (X) of the t-SNE algorithm for the different response/treatment pairs. B) shows the same comparison for the second dimension (Y) of the t-SNE algorithm. The vertical scale values are mean centered.

5.4.3 Pathway Analysis Highlights Conserved Genes Indicative of Non-response Pan-treatment

t-SNE representation shows there is a significant clustering of the non-responsive patients pan-treatment using the pairwise fold change values of the pre and on-treatment samples as input. In order to establish whether there was a significant biological underpinning for the gene lists, pathway analysis were performed and the results can be seen in Figure 5.4.3, page 143. This result is suggestive of common expression profile changes in patients with no response to treatment not present in either of the responsive classes. Combining the overlap of prominently enriched genes from gene set enrichment score analysis via GSEA and CAMERA filtered gene lists revealed significant enrichment of the mTOR signalling and PI3K-AKT signalling KEGG pathways as well as Breast cancer associated, p53 signalling, TNF, Insulin, ErbB and MAPK signalling pathways. In addition an elbow diagram was plotted to identify the most important subset of genes which contribute most heavily to the pathway analysis results. This type of diagram seeks to highlight an inflection point which shows a change in the behaviour of the samples, which becomes apparent as an “elbow” bend in the plotted data. Pathway analysis of this subset of genes was performed and the results brought forward for further analysis to see if the reduced gene space would offer a more concise view. Analysis of the surrounding literature of non-responsive and TNBC breast cancer reveals mTOR and kinase inhibitors to be of special interest to the treatment of these diseases. A further examination of the gene lists was undertaken to examine the gene specific changes that contribute to this result. There were significant differences in several of the genes constituting the differential feature list used for the pathway analysis, Figure 5.4.4, the two genes chosen for this diagram are MDM2 and RBL2, they are representative of the behaviour of the selected genes, but were selected in the subsample of two maintain visual clarity of the diagram. This supports the pathway analysis and t-SNE diagrams that the pre- and on-treatment expression level change values are capturing patient specific response.

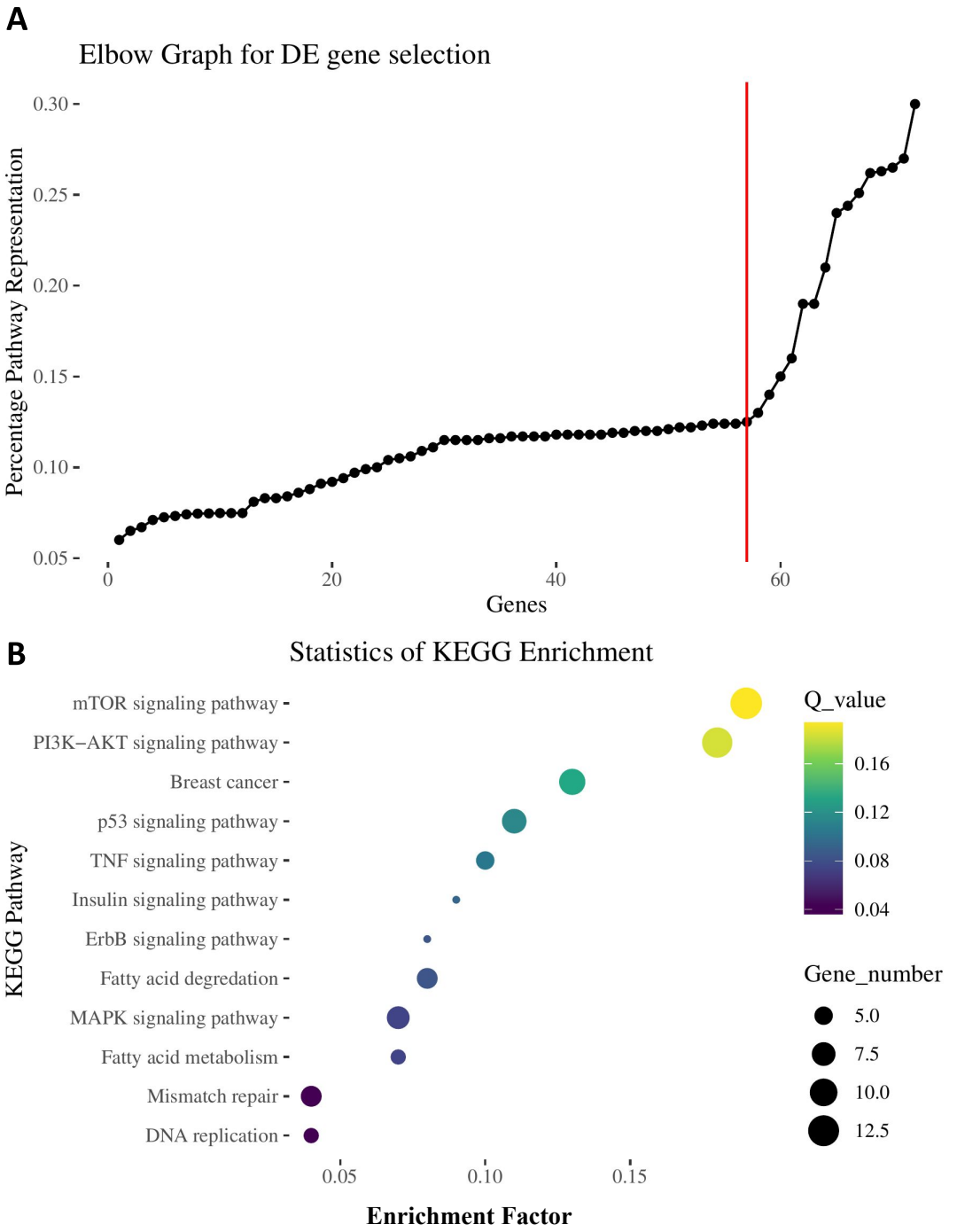


Figure 5.4.3: KEGG Pathways Enriched in Non-responding Tumour Samples. A) This "Elbow" graph is a dot plot showing the percentage of KEGG pathways identified as being enriched in the non-responders compared to the responders. This shows that a small portion of genes are present in an large percentage the KEGG pathways. B) Shows the KEGG pathway enrichment scores that the genes selected by the vertical line in A select for. These are pathways enriched in the non-responders compared to the responsive patients when measuring the fold change values, looking in both treatments.

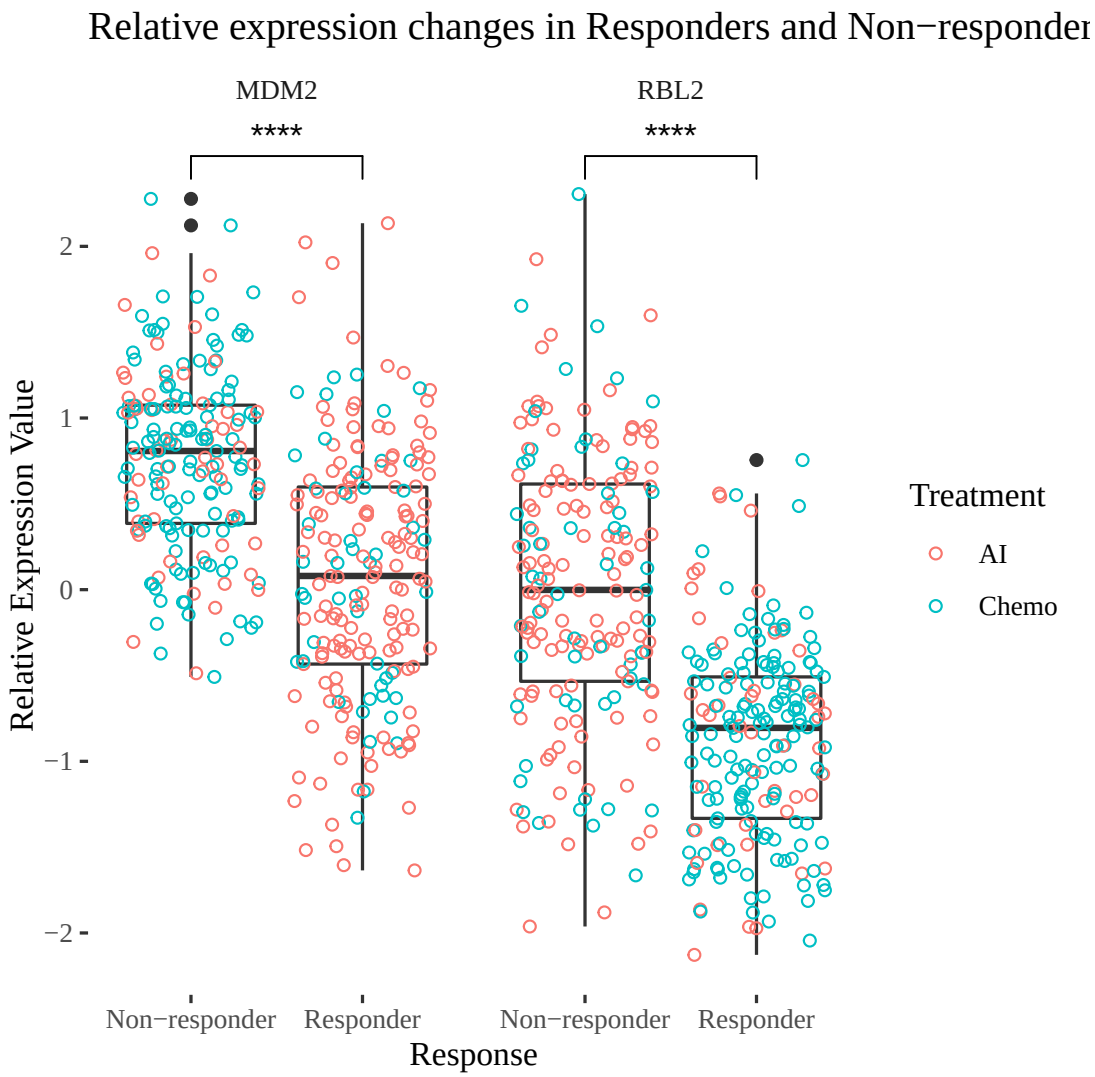


Figure 5.4.4: Differentially Expressed Genes Show Significant Differences Between Response Classes MDM2 shows an increase in expression level on-treatment in the non-responding patients, while a proliferation associated marker RBL2 shows a decrease on treatment in the responders, but a steady level in the non-responders. These are two of the genes that see significant differential expression between the overall responsive classes and all labeled non-responders. There is no significant difference in the fold change for these genes if the data is faceted on treatment type.

5.4.4 Identification of Non-responsive Patients from Pairwise Delta Expression Values

In order to determine if the aggregated differential gene lists from the pre- to on-treatment matched patient expression levels changes could be used for the accurate prediction of response, two models were trained to predict response independent of treatment. 823 Samples with available response status representing 365 patients were subset into an 80:20 training and test split and two models were trained on the response classes, regardless of treatment paradigm, using the differential gene lists to create a predictive model with ten fold cross validation. As the classes were well balanced, 385 non-responders, 438 responder (skew = 1.137) accuracy could have been used as the test metric, but to ensure fair testing the precision, recall, and F1 scores are reported in Table 5.2, page 145. F1 is a measure of accuracy that takes into account both the precision and recall for a model, and is calculated as $F_1 = \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$, this metric was chosen over AUC/ROC as it is less sensitive should a subset of validation data have a class imbalance. SVM using a radial boosting function was trained to find the boundary of the response class areas and had an F1 score of 0.91 in training and 0.89 in testing on our split data. In addition, a neural network was constructed using Tensorflow and Keras with two dense layers and a 0.1 fractional dropout layer to avoid over fitting, which resulted in a training accuracy of 0.97 and a test validation of 0.92. These models were both highly accurate at detecting overall response from these pre- and on-treatment matched expression level values, clearly indicating that the relative changes seen on-treatment are important for stratifying risk and predicting patient response.

Model	Precision	Recall	F1	SVM	P.Responder	P.Non-Responder
SVM	0.88	0.9	0.89	Responder	180	24
NN	0.92	0.93	0.92	Non-Responder	20	161
				NN	P.Responder	P.Non-Responder
				Responder	190	16
				Non-Responder	14	165

Table 5.2: Machine Learning Models Report High Accuracy for Predicting Response From Matched Sequential Samples. This table presents the confusion matrix that results from each ML classification technique, SVM and NN. The columns of each matrix represent the "Predicted" status (P.Responder, P.Non-Responder), and the rows the ground truth for the model to compare its performance to. SVM and NN's were built and trained on an 80/20 split to identify response. The precision, recall, and F1 are listed in the table.

This analysis was also repeated by removing one of the five data sets to create a 4:1 split of available training data. The models were then trained on samples

from the four aggregate datasets and validated on the withheld samples. This extra cross validation was performed on all five possible combinations of training/validation pairs and the results are tabulated in Table 5.3, page 146. The performance of the SVM and NN models under these conditions ranged from an F1 score of 0.88-0.93 in the NN and 0.85-0.91 in the SVM. The affect of this alternative testing strategy on the classifier performance shows that these models are affected by the training sample size as the lowest performance (0.85 F1 in the SVM, 0.88 in the NN) in both models was when the largest data set was removed from the training cohort (E1, 372 samples) and the highest performance (0.91 F1 SVM, 0.93 F1 NN) was when the smallest data set (C1, 95 samples) was removed. Comparable analysis of training on one treatment type and validating in the other was not performed as this would have exposed new biases into the training models. The chemo therapy cohorts were of markedly worse prognostic grade (endocrine treated cohorts mean grade at diagnosis 2.11, chemotherapy, 2.6) as well as the sample imbalance between the two cohorts would have created models of variable robustness.

Removed Dataset	NN	SVM
	F1	F1
E1	0.88	0.85
E2	0.90	0.87
E3	0.90	0.88
C1	0.93	0.91
C2	0.91	0.90

Table 5.3: Cross Dataset Validation Accuracy Values Comparing Different Subsets of Data. This table shows the balanced F1 accuracy scores for the testing validation of the pre-trained NN and SVM models trained on four datasets, then tested on the withheld samples. E1 is the largest dataset and the affect on the accuracy is noticeable when it is removed from the training cohort. Conversely, training on the combined data except C1, the smallest dataset, shows a small improvement in accuracy in validation.

5.4.5 Breast Cancer Transformation and Classification

The possibility of transforming the tabular data to images as a means of preprocessing was also investigated. The pairwise correlations of matching samples pre- and on-treatment samples were calculated and plotted as heatmaps which were saved and labelled as responder or non-responder. A ResNet model was fine tuned on a randomly sampled 80:20 split and then validated on the hold out data. This resulted in an F1 accuracy of 0.90 which is a marginal improvement of 0.01 over the deep neural net work alone. Representative heatmap samples

are included in 5.4.5 147.

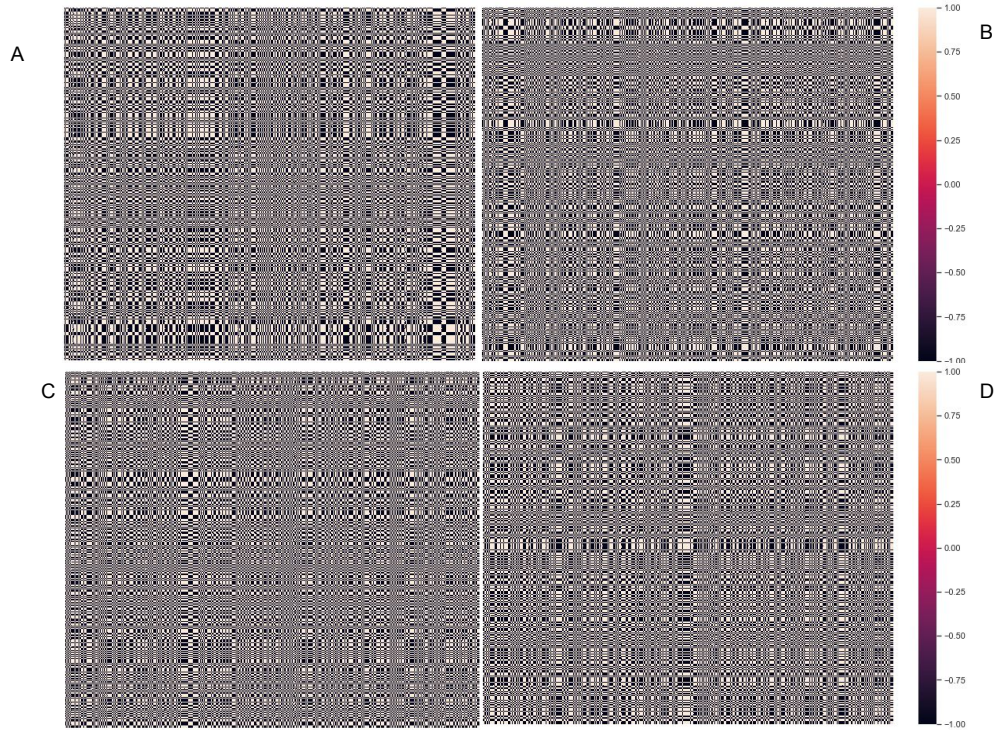


Figure 5.4.5: Breast Cancer Treatment Response Classification of Correlation Heatmaps. Plots A, B, C and D are representative examples of the correlation heatmaps that result from the pre- to on-treatment expression value changes. By training on these files with associated known response a deep convolutional neural network was trained to classify response in a parallel method to the tabular classification approach.

5.5 Discussion

The results of this analysis reveal a few potentially important and suggestive outcomes as to the net change experienced as part of normal neoadjuvant therapy of breast cancer. Previous work has highlighted that there are important genes differentially expressed in response to chemotherapy[273,274] and endocrine therapy[275] and that it is clear that these lists are not identical. Common threads are a decrease in the level of proliferation markers[276,277] and apoptosis in both paradigms. However, the initial analysis of the changes seen in chemotherapy are mirrored, to a lesser extent in the endocrine cohorts suggesting some commonality in the overall expression profile of antimitotic therapy, regardless of target.

Possibly of more clinical interest is the perceived overlap in the trajectory

of gene expression for non-responsive patients in both treatment type cohorts. While the pathways of response differ between endocrine treated and CT, the pathways of non-response show greater similarity and unsupervised analysis reveals them to appear to be the same sub-population. This is potentially of great interest to novel drug discovery, as this could potentially lead to new targeted therapies. This statement is emboldened by the results of the differential gene expression analysis, pathway enrichment, and gene set scoring methods, which show conserved similarity in the pathways between total response and non-responsive patients. KEGG analysis showed significant enrichment of the PI3K and mTOR pathways; two promising contemporary therapies to more resistant tumours.[278,278,279]

Lastly, models were built using standard statistical machine learning and neural network approaches for the identification of response from the on-treatment expression vectors with marked success in testing (SVM: 89%, NN: 92%). This is highly concordant with the accuracy observed in previous studies of applied ML to predicting risk and recurrence from clinical and pathological features (decision trees, CNN and SVM are 0.936, 0.947 and 0.957 accuracy respectively)[280] and imaging data.[281] This is further support for the work detailed in Chapter 2, which began to enumerate the value of treatment samples for this precise purpose. These results are of imminent interest to the body of work surrounding this text in the area of biomarker identification and risk stratification.

5.6 Conclusion

Meta-analysis of multiple pre- and on-treatment neoadjuvantly treated datasets has revealed several key insights into the value of sequential sampling and the applicability of pan-treatment markers for response in primary breast cancer. *En masse* some of the changes seen in one treatment are seen in the others and the pairwise differences observed in different response classes can be utilized to predict response from the pairwise fold change values captured by sequential sampling. Patients that respond to chemotherapy and patients that respond to endocrine therapy have different relative gene expression profiles, but the non-responders to both therapies show strong overlap in relative expression from the sequential samples. Lastly, supervised machine learning algorithms trained on the sequential fold change values were highly accurate for the prediction of response in both the neoadjuvant chemotherapy and endocrine settings.

6 | Perspectives, Discussion, and Conclusion

An examination of The National Center for Biotechnology Information reveals that there are over 505,000 papers returned for the query *Breast Cancer*, as of October 2019. The prospect of adding new and novel work to this body of literature poses an enormous task and the contribution of this thesis to that corpus must be considered fully. While there are clear and distinct areas of prominent research, such as biomarker discovery, new targets for therapy, and identification of new actionable subtypes, my own analysis of the 100,000 most cited papers reveals many more distinct groupings in this domain, see Figure 6.0.1 on Page 150. Finding my own niche in this extensive body of research helped to define the parameters of my work and lead me to the results and output I have produced during my PhD. In this thesis, I seek to fully explain the context of my research, its impact, and what this work may lead to in the future.

6.1 Discussion

This thesis work is broken down into four discrete results chapters each composed of modules of my work over the past four years but which in aggregate represent a larger and more valuable message. Here I will discuss the interplay between the chapters and begin to introduce the conclusions which should be drawn from this body of work. I will highlight the poignant results of each area of my thesis, the setting and surrounding literature, and potential avenues of further research. A fundamental problem in the treatment of breast cancer is the heterogeneity of the disease[282]. Breast cancer doesn't represent one disease but many, each with unique pathological and treatment options.[283] Indeed, even among a constituency of similar tumours, patient variation can still occlude matched tumour characteristics. In this regard, pairwise matched samples taken during neoadjuvant therapy help to reduce the variance of the data set, and allow for a remarkably granular view of patient risk on a personal level.

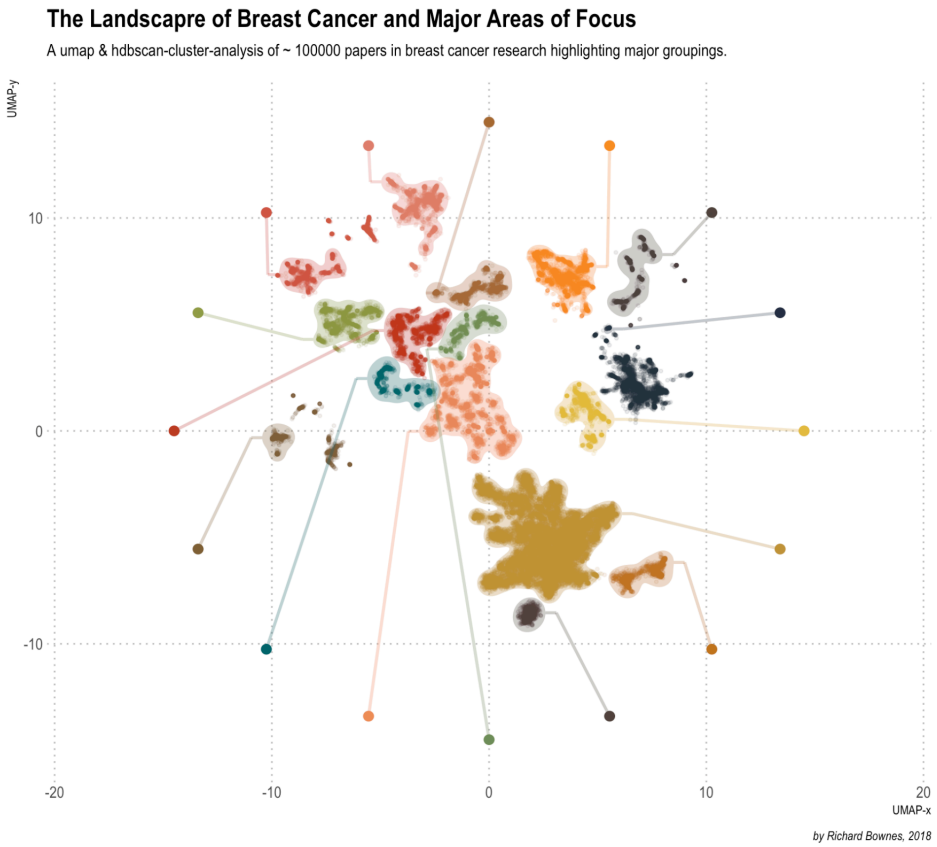


Figure 6.0.1: UMAP reduced representation of breast cancer research. 100,000 breast cancer research papers were grouped by filtered content, authorship and citations to reveal distinct areas of research.

This is due to the fact that the inter-patient differences may be removed when considering only the pairwise gene expression values from sequential samples within each patient.[284]

In addition to capturing the response of each patient from the mid-chemo time point of a sequentially sampled cohort, it was also able to forecast the survival of patients in both studies in a highly significant manner. This clearly demonstrates that a marker generated on the delta values of gene expression is able to capture response class specific changes to chemotherapy that are uniform among patients and cohorts at least partially due to the fact that the signal to noise ratio present in the training samples is boosted by removing the inter-patient variations. It should be noted that biomarker identification is a topic of much scrutiny. Single novel biomarkers may be subject to noise and variation and highly dependent on the subset of patients used to identify the signature.[285] There is also evidence that random biomarkers may be predictive as many are associated with proliferation, a positive indicator of response globally[286] and that realistic datasets for true identification of biomarkers would require massive sample sizes,[287] the goal of chapter 4. An important take away is more the value of the on-treatment samples and the extra information held in the pairwise gene expression change values. In addition, it is strongly suggested that the on-treatment samples may be of increased prognostic value from a subtyping and stratification perspective, as the on-treatment samples had a demonstrably differential expression profile and a greater ability to "correctly" reflect the actual risk posed to each patient based on calculated metrics of prognostic risk (MammaPrint, rorS). This is highly novel work, as there is a paucity of preceding work in biomarker generation from on-treatment samples, the majority of the accessible results have been included in this manuscript already, primarily from Sims *et al.*. Markers of proliferation have been previously studied for their ability to stratify patient risk, but this work is tangential, not identical to the methods presented here. This does raise the question of what additional biomarkers can be identified from these matched patient samples, especially when larger uniform datasets become available.

Matched patient sequential samples of primary breast cancer are not the only additional samples taken in the routine clinical and pathological appraisal of BC. Tumour positive nodes are also frequently examined as a categorical risk indicator for future recurrences and metastasis. However, as I explore in chapter three, the same pairwise relative expression differences can be utilized in this new setting to improve the capture rate of patient recurrence. The gene expression level changes that facilitate the escape of the primary tumour from it's tu-

mour bed and to infiltrate beyond are over represented in the tumour profile of the sampled lymph node. This differential pattern of expression provides a more accurate sampling of the patient specific risk posed by their metastatic disease and can potentially be useful in improving the treatment of these especially high risk patients. In this regard the additional sampling not for the purpose of observing the on-treatment gene expression changes, but to identify the tumour's evolution beyond the bounds of it's original tumour bed is what provides the new and requisite information for this analysis. Pairwise concordance of primary BC and metastatic disease has previously been examined for mutational differences[288,289] to understand the progression of disease and for identifying new actionable targets[290] and modern studies have identified gene expression profiles between the two.[291] However this is the first time that such a comparison has been used to better estimate the risk posed by the primary diagnosis. The data presented in this cohort is also the largest matched primary and lymph node dataset currently available with follow up and on-treatment paired biopsies. The availability of contemporary datasets like the Lawler trial have greatly boosted the validity of the claims made from this analysis as the same patterns of expression and relative risk are clearly present in this parallel cohort.

From a high level the premise of more samples, better results seems intuitive. More powerful statistical analysis is possible, which means results can be better supported and the impact of findings fundamentally of increased integrity.[292] However, it isn't the static increase in dimensionality that provides the importance of these samples, partially as the patient number is not inflated, but the relative changes that grant new insight and understanding to breast cancer treatment and patient options. In the first instance this new information is gained by observing temporal changes in differential expression while minimizing the noise present in the comparison of the data. This has been previously examined to some extent where time scale analysis of gene expression reveals new pathway enrichment information from these fold changes.[293] These pairwise on-treatment difference have also been suggestive of helping to identify the mechanistic drivers for aberrant molecular pathways for patient and tumour characterization.[294] In the latter, new value is added by probing biopsies to see a facet of the disease that is not present in the primary tumour body, but more pronounced in the metastatic reaches of the disease. Patient matched and sequential samples are rare however, and no unified resource or repository is available to freely probe on-treatment samples for further biomarker identification or validation testing. The creation of MetaMatchedBreast to improve the analytical prospects was inspired by this precise lack of resources. This standalone R package wraps up expression sets,

with further clinical information and considerable built-in tools for analysis and biomarker comparison as well as cross study and treatment type response and survivorship analysis. This will hopefully provide a useful functionality to future researchers in this field of neoadjuvant breast cancer research.

In my fourth chapter I prepare two references to compare the performance of quantile normalization with ComBat batch correction with additive covariates of integration. First, pre-treatment samples only, second, using no integration methods and just retaining similar features. While integration of the pre-treatment alone sample is supported by literary evidence as a viable means of integration,[216,226,240,241] matched pre and on-treatment samples have never been integrated as far as I am able to determine, hence the need to establish base lines for performance. Integration of the underlying expression values was exhaustively shown to be inadvisable from my results, but this lead to the creation of a new bioinformatics resource for the examination of these matched patient samples. This is a truly novel new tool for analysis as to date there are no available such tools in the neoadjuvant or sequentially sampled space. Other existing tools for meta analysis have focused exclusively on the aggregation and analysis of the most available breast cancer data; pre-treatment only samples. MetaMatchedBreast fills a niche not currently serviced by the existing tools for breast cancer meta-analysis.[73,211,212] This leads to the distinct prospect of future improvements in the understanding of the mechanistic and pathway related changed in response to therapy,

MetaMatchedBreast refocuses instead on the most important features of this data; the pairwise expression level changes. Here we avoid bringing together the multiple distributions, and sub distributions of data present in each dataset. Instead, by normalizing the values consistently and retaining the delta values between each time point, the same-scale pre and on-treatment changes can be compared across multiple studies. This facilitates the increased statistical significance that this type of analysis has previously lacked, while emphasizing the very features of this data that make them unique and additionally valuable. This is possibly the most appropriate way of combining this type of data as it has been suggested that integrative analysis falsely conveys confidence on results from aggregated data.[295] Analysis of this data as pairwise pre and on-treatment values immediately yielded results. Clustering analysis of patients with known response highlighted that responsive chemotherapy and aromatase inhibitor therapy have significantly different patterns of expression change. However, there was dramatic overlap of the non-responsive patients from both groups. Upon performing pathway analysis and enrichment scoring between the groups of

responsive and non-responsive patients, genes present in the mTOR and PI3K pathways were significantly over expressed in the non-responders. While not definitive, this is at least suggestive of common escape pathways in BC that may be utilized for future drug discovery. These pathways highlight the resolution of known and important pathways in breast cancer, mTOR[296–298] and PI3K[299–301] that were over represented in the non-responsive samples and may represent an exploitable future avenue for improving the outcome of these patients. This work is similar in some regards to other network analysis of BC that has previously identified similar such pathways in non-responsive or resistant patients[302]. However, the unique nature of this pairwise difference patient data may help to elucidate further insights into the causes of non-response.

There are several other confounding factors which have varying degrees of importance to the outcome of this thesis but which need to be considered fully for subsequent work. Cell type heterogeneity was never explicitly examined in the needle-biopsied samples. This means that there are fundamental assumptions in this work. Of primary concern is that the tumour samples provided by the clinicians and examined by pathologists were 'tumour'. As to my knowledge no heterogeneity scores were calculated or estimated I can not say with certainty that this is the case, and that perhaps with a full cadre of these results would have drawn some conclusions differently, or performed analysis in ways that would be more tolerant of non-homogeneous sampling. This is especially important when considering that in a recent review it was shown that in multiply sampled breast cancer from the same patient there was remarkable intra-tumour geographic heterogeneity[303].

In a similar vein, in some contexts fresh frozen and FFPE samples were included in the analysis. While all possible care was taken in the handling of these results, and cross-platform-normalization was performed to unify these dataset it should still be noted that FFPE samples show uniformly reduced differential expression when compared with fresh frozen.[304] This is important when comparing results as the relative expression of key genes related to breast cancer risk assessment and analysis will be less pronounced in these cohorts, potentially reducing the power of the entire study.

A major theme of this thesis is to determine whether the existing diagnostic/pre-treatment methods like PAM50 work more effectively on samples that have seen chemo/endocrine therapy. The assumption here is that once tumour cells have been treated, even for as short a period as two weeks, the tumour response will become manifest in the expression profile in a more

pronounced way than pre-treatment and actual, categorical, risk of the patient will be more accurately captured by these tests. While the results proposed in this thesis indicate that this is indeed the case, the overall number of samples in these studies is low and testing on greater cohort sizes would be preferable to eliminate the chance of this being a statistical artefact. Alternatives to this would include developing new tests, like the gene marker proposed in results chapter one, designed for use exclusively on the treated samples, or simply improving the diagnostic accuracy of the existing tests. There does exist a serious concern as to the use of these methods. As they take expression matrix values as the input these values can range due to the platform of analysis, the means of normalization, non-homogeneous normalization methods and as it turns out batch correction techniques.

A final assumption, one that was made out of necessity due to the limitations of sample size and the different reporting of response between datasets is that all responders look the same, and all non-responders look the same, from a gene expression stand point. By defining this from the onset we are making the strong assumption that they can be grouped, and neglecting to include partial response as a category or allowing for the possibility that there are alternate escape pathways to response than than what is present in the majority. This assumption was made due to the relatively small size of the data, and the fact that pathological complete response was the only method for assigning response on mass that was available across each study. This is an unfortunate product of circumstance, but one that with more time could be rectified by the inclusion of further studies. In order to try and test for the possibility of sub groups of responders, multiple iterative clustering approaches were performed in, but likely as a result of the aforementioned limits on dataset size there was no significant sug-groupings of samples within each response category.

A last piece for consideration is the crisis of reproducibility in academia and how bioinformatics should be leading the charge towards better result replication, but why we aren't. In an attempt to be as forthcoming and transparent as possible I have hosted all of my scripts and publicly accessible data on a GitLab page, ideally facilitating the easy sharing a reproduction of my results with other researchers. However, outside of these containerised solutions it is worth noting that at a systematic level the results of this thesis, and many others like it, could invariably fail to work across data centers, across platforms, across different treatment types and pathologies. This is down to two factors, the massive heterogeneity of disease meaning that biomarkers are usually developed for a specific subset of disease, something this thesis sought to address, and the com-

plete lack of a standard when it comes to the collection, storage, analysis and processing of patient samples. Fundamentally, one of the most important reasons biomarkers fail is down to this lack of standardization, essentially creating microcosms of treatment/processing pairs in which solutions are viable, but not outwardly extensible.

6.2 Conclusion

This thesis fundamentally set out to prove or disprove the following hypothesis:

> *If patient-matched on-treatment or lymph node positive samples are informative for the classification and characterization of breast cancer, then they should facilitate improved differential and statistical analysis of breast cancer.*

In this regard this thesis has been successful in providing evidence for the additional information available through the pairwise analysis of breast cancer from the perspective of these samples. From my work in this thesis, I have shown clearly that the on-treatment samples can be utilized for the creation of novel and highly accurate biomarkers for the prediction of response to chemotherapy and as a strong positive indicator of survival. In addition, these samples, and other pairwise tissue sampling of metastatic lymph node tissue are more informative for the use of existing risk scoring and profiling tests for the prognostic stratification of patients. Lastly, that in aggregate, the pairwise differences between the pre- and on-treatment samples can be a powerful tool for pathway analysis to reveal new insights into the common mechanisms of tumour response escape and identifying important features that define these non-responsive patients.

Appendix: Published papers

Bownes et al. Breast Cancer Research (2019) 21:73
https://doi.org/10.1186/s13058-019-1159-3

Breast Cancer Research

RESEARCH ARTICLE Open Access

On-treatment biomarkers can improve prediction of response to neoadjuvant chemotherapy in breast cancer

Richard J. Bownes[1], Arran K. Turnbull[1], Carlos Martinez-Perez[1], David A. Cameron[1,2], Andrew H. Sims[1*] and Olga Oikonomidou[1,2]

Abstract

Background: Neoadjuvant chemotherapy is increasingly given preoperatively to shrink breast tumours prior to surgery. This approach also provides the opportunity to study the molecular changes associated with treatment and evaluate whether on-treatment sequential samples can improve response and outcome predictions over diagnostic or excision samples alone.

Methods: This study included a total of 97 samples from a cohort of 50 women (aged 29–76, with 46% ER+ and 20% HER2+ tumours) with primary operable breast cancer who had been treated with neoadjuvant chemotherapy. Biopsies were taken at diagnosis, at 2 weeks on-treatment, mid-chemotherapy, and at resection. Fresh frozen samples were sequenced with Ion AmpliSeq Transcriptome yielding expression values for 12,635 genes. Differential expression analysis was performed across 16 patients with a complete pathological response (pCR) and 34 non-pCR patients, and over treatment time to identify significantly differentially expressed genes, pathways, and markers indicative of response status. Prediction accuracy was compared with estimations of established gene signatures, for this dataset and validated using data from the I-SPY 1 Trial.

Results: Although changes upon treatment are largely similar between the two cohorts, very few genes were found to be consistently different between responders and non-responders, making the prediction of response difficult. AAGAB was identified as a novel potential on-treatment biomarker for pathological complete response, with an accuracy of 100% in the NEO training dataset and 78% accuracy in the I-SPY 1 testing dataset. AAGAB levels on-treatment were also significantly predictive of outcome ($p = 0.048$, $p = 0.0036$) in both cohorts. This single gene on-treatment biomarker had greater predictive accuracy than established prognostic tests, Mammaprint and PAM50 risk of recurrence score, although interestingly, both of these latter tests performed better in the on-treatment rather than the accepted pre-treatment setting.

Conclusion: Changes in gene expression measured in sequential samples from breast cancer patients receiving neoadjuvant chemotherapy resulted in the identification of a potentially novel on-treatment biomarker and suggest that established prognostic tests may have greater prediction accuracy on than before treatment. These results support the potential use and further evaluation of on-treatment testing in breast cancer to improve the accuracy of tumour response prediction.

Keywords: Breast cancer, Chemotherapy, Gene expression, Response, Outcome, Predict, Neoadjuvant, Biomarker

* Correspondence: andrew.sims@ed.ac.uk
[1]University of Edinburgh Cancer Research UK Centre, MRC Institute of Genetics and Molecular Medicine, Edinburgh, UK
Full list of author information is available at the end of the article

Introduction

Chemotherapy is among the most common effective treatments for breast cancer, alongside radiotherapy, hormone therapy, and targeted treatments. Neoadjuvant chemotherapy is given prior to surgery with the aim to reduce the tumour burden and to provide early information on the response to treatment [1]. Studies have shown patients with tumours that have a pathological complete response (pCR) following neoadjuvant chemotherapy are much less likely to recur than those in women with residual disease [2]. Neoadjuvant chemotherapy is now considered as the standard of care in breast cancer and has seen a rise in recent years with data from powered studies suggesting that the pathological complete response achieved following neoadjuvant chemotherapy might be a surrogate of good prognosis [3]. A recent meta-analysis also showed significant tumour response and an increase in the rate of breast-conserving surgery following NACT with good rates of long-term local recurrence (5.5% vs. 15.9% adjuvant chemotherapy), however with an increase in the rate of short-term local relapses (1.35 RR 0–4 years, 1.53 RR 5–9 years) [4].

Neoadjuvant treatment provides a "window of opportunity" (Fig. 1a), where sequential sampling of a tumour enables observation of the changes that occur in response to treatment to be measured and considered in the context of response and outcome [5]. Neoadjuvant therapy studies and pre-surgical treatments allow for a unique in vivo analysis of tumour treatment response [6], as well as the possibility of predicting the response to treatment earlier in the treatment [5]. It has been suggested that on-treatment biomarkers may be superior to those measured before exposure to treatment [3, 7]. On-treatment information has already been shown to be informative for the accurate prediction of response to endocrine therapy [8]. Here, it was found that patients with elevated Ki67 levels (higher than 10%) at 2 or 4 weeks exhibited resistance to endocrine therapy and were triaged to neoadjuvant chemotherapy [8]. We have also demonstrated the potential of on-treatment biomarkers by developing a four-gene signature which combined pre-treatment expression levels or two biomarkers (IL6ST and NGFRAP1) with patient-matched 2-week on-treatment expression levels of two proliferation

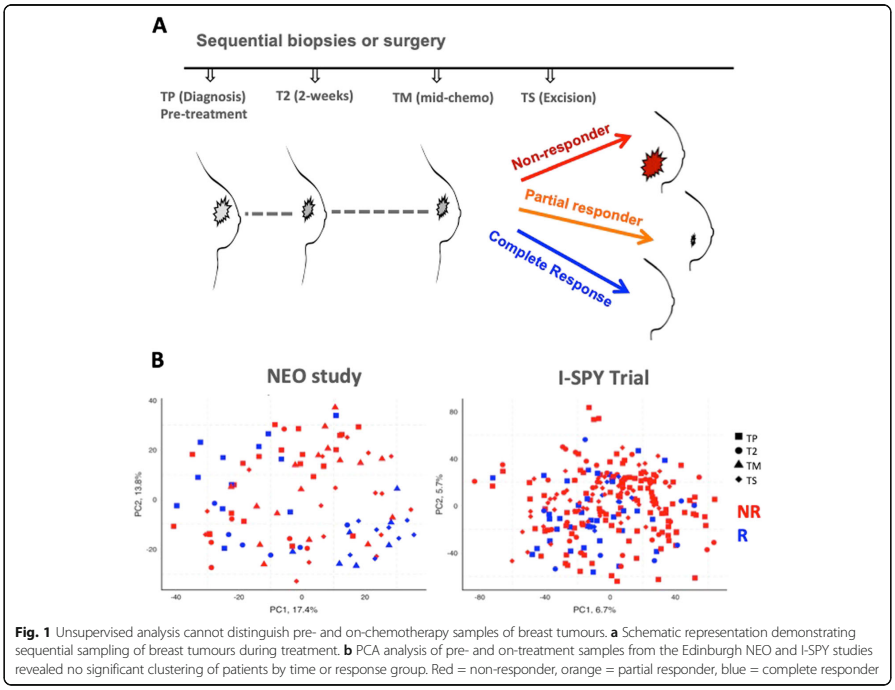


Fig. 1 Unsupervised analysis cannot distinguish pre- and on-chemotherapy samples of breast tumours. **a** Schematic representation demonstrating sequential sampling of breast tumours during treatment. **b** PCA analysis of pre- and on-treatment samples from the Edinburgh NEO and I-SPY studies revealed no significant clustering of patients by time or response group. Red = non-responder, orange = partial responder, blue = complete responder

markers (ASPM, MCM4) to accurately predict the response to endocrine therapy in a blinded independent validation set [7].

Gene expression-based studies of neoadjuvant chemotherapy treatment to date have largely been limited to studying the association of pre-treatment samples with pathological response [9, 10]. Patient-matched sequential sampling gene expression studies have been previously attempted; however, they have not evaluated the predictive capacity or proposed new on-treatment predictive biomarkers [11–13].

In this study, we present the largest sequentially sampled patient-matched analysis of neoadjuvant chemotherapy-treated breast cancer tumours to evaluate whether on-treatment biomarkers can improve the accuracy of predicting response before resection. Numbers of patients with sequential breast tumour samples are limited, but we compare and validate our results with the data from the I-SPY 1 Trial.

Materials and methods

Patients, response criteria, and samples

The NEO study consists of 50 breast cancer patients with sequentially sampled biopsies at four time points, pre-treatment (PT, 34 samples), 2 weeks on treatment (T2, 12 samples), mid-chemo (TM, 23 samples), and at surgical resection (TS, 24 samples) with three clinically defined response statuses: complete responders (pCR by resection), good responders (tumour volume reduction, but lack of pCR), and non-responders (progressive disease or small tumour volume changes on treatment). Patients were of mixed histological grade and HER2 status; ages ranged from 29 to 76. Patients were primarily treated with 3 cycles of FEC and docetaxel with Herceptin where appropriate. Three patients received paclitaxel, one patient received additional carboplatin, one patient received Epi-cyclophosphamide and paclitaxel, and one patient received docetaxel and cyclophosphamide. Eligible patients were women with histologically confirmed invasive breast tumours and with no evidence of distant metastatic disease, no prior history of malignancy, and fit enough to receive chemotherapy in the opinion of the responsible clinician irrespective of age. All cases were discussed at the breast MDM in Edinburgh Breast Unit at the Western General Hospital, and consensus from this meeting was to be treated with neoadjuvant chemotherapy.

Core needle (16-gauge) biopsies were taken from the primary breast tumours before treatment (PT) and between 10 and 14 days after the first dose (T2) of chemotherapy. A third sample was taken at the mid-chemotherapy point day 20–21 (TM), and finally, a core biopsy was taken from the excision specimen (TS) after it has been removed prior to submission to pathology. Fixed and frozen samples of normal and tumour tissue were collected from all specimens.

Gene expression profiling

RNA extraction was performed via Ribo0-RNAseq, and whole transcriptome sequencing was performed with Life Sciences Ion AmpliSeq™ Transcriptome Human Gene Expression Kit. This generated greater than 8 M reads per sample with an average of more than 90% valid reads for 12,365 targeted genes. Most analyses were performed in R (http://www.r-project.org) using packages available through CRAN (http://cran.r-project.org/) and Bioconductor (http://www.bioconductor.org/). Outside of the R environment, the stand-alone application Multiple Experiment Viewer (http://mev.tm4.org/) was utilised for pairwise ranked product feature selection, and DAVID (https://david.ncifcrf.gov/) was used for pathway identification. Additionally, the python package scikit-learn [14] was used for unsupervised clustering analysis. Ninety-seven samples were analysed over 13 AmpliSeq chips, but no systematic batch effects were evident and no batch correction was performed within the training data. Gene expression data for the NEO study has been made publicly available at the NCBI GEO data repository under accession GSE122630.

The I-SPY 1 Trial is composed of patients with invasive breast cancer > 3 cm, or at least one tumour-positive axillary lymph node [11]. Patients were treated with an anthracycline-based chemotherapy followed by taxanes [11]. Samples were normalised and corrected for background red/green signal; Bioconductor R packages marray and limma [15] were used to this end. From the original 221 patients, only 36 had matching pre- and on-treatment samples, and 39 had matching biopsy and excision samples; pathological complete response was used for response criteria. Pairwise gene expression was handled with SAM and follow-up analysis with Ingenuity Pathway Analysis from QIAGEN Bioinformatics. I-SPY 1 Trial data is hosted at NCBI GEO under accession GSE32603 [11].

Statistical analysis methods

Principal component analysis (PCA) was performed on unsupervised gene lists to reduce dimensionality and visualise differences in response at all times and to identify present differences between patient treatment statuses. Local Fisher discriminant analysis (LFDA) [16] was used at each time point to determine if the response groups could be distinguished with treatment time with a semi-supervised clustering approach, concurrently with class advised *K*-means clustering. LFDA is a form of supervised dimensionality reduction that maximises between-class scattering and minimises within class scatter, and is a refined version of normal Fisher discriminant analysis [16];

this exploratory analysis was used in order to visualise comparative differences in treatment time, not as a means of feature selection. Pair-wise significance analysis of microarrays [17] using the *siggenes* package in R was used to consider the consistency of differentially expressed genes due to treatment in the sequential patient-matched samples. Rank Product analysis was used to identify differentially expressed genes between response classes at each time point. Successive levels of standard p value (0.05, 0.01, 0.001), without correction for multiple testing, were used in order to determine the number of differentially expressed genes, and at lower p values which the time points had the most strongly differentiating genes. Significance analysis of microarrays was also performed using varying false discovery rates (1%, 5%, 10%) to try to identify common differentially expressed genes between responders and non-responders across both datasets at each time point. Gene score enrichment analysis was used to validate the time point selection by looking for the highest number of enriched pathways. The gene list from the most differential time point (TM) using the NEO dataset was extracted and used in a random forest model (10,000 trees, m-try as the square root of the feature number) using pCR status as the class label (clinician-identified pCR and non-pCR). The most deterministic genes for class prediction were fed into a classification and regression tree in order to produce a maximally reduced and repeatable model; this methodology is further described by Turnbull et al. [7]. The CART decision tree was applied to the NEO dataset for training and tested in the independent I-SPY 1 dataset using the same cut-points determined by mean-centring the datasets. This protocol was repeated using the gene list from the pre-treatment only samples, using the same p values and tree configurations for selection. Survival analysis was performed at different time points using the log-rank test. Intrinsic subtypes, Mammaprint, and risk or relapse scores were estimated from the gene expression data using the *GeneFu* R package [18].

Results

Gene expression differences between responding and non-responding breast cancer tumours treated with chemotherapy are subtle and time dependent

Unsupervised principal component analysis was first used to assess whether sequential patient-matched samples from patients receiving chemotherapy (Fig. 1b) would cluster by time point or response status. There was no significant grouping of patients according to sampling time: pre, early, or later after chemotherapy in either the NEO or I-SPY 1 studies (Fig. 1b). There were no significant differences between the two cohorts in terms of age, grade hormone receptor, and HER2 status, and the subset of patients with mid-chemo samples was not significantly different from the whole NEO cohort (Table 1). Patient-matched samples enable the pairwise analysis to look for consistent changes in the gene expression during treatment. Pairwise significance analysis of microarray analysis using a 10% false discovery rate (FDR) identified a relatively small proportion of overlapping upregulated (5%) and downregulated (4%) genes between the two studies. However, genes that were increased or decreased in response to treatment in one study were also clearly and consistently increased or decreased in the other study (Additional file 1: Figure S1A), further suggesting it would be difficult to discriminate responders from non-responders. Indeed, there was no clustering by response status before or during treatment (Additional file 1: Figure S1B). These results likely reflect the considerable inter-patient differences being substantially larger and more significant than the subtler commonalities in gene expression of a particular time point or response class of each tumour. More encouragingly, semi-supervised LFDA of each time point

Table 1 Summary of patient characteristics for the NEO study and I-SPY validation set

Characteristics	NEO cohort (50)	NEO cohort PT-TM pairs (23)	p value	I-SPY 1 PT-T2 pairs (36)	p value
Median age at diagnosis	50.8	50.1	0.8	47	
Tumour grade			0.52		0.58
1	0 (0%)	0 (0%)		1 (3%)	
2	22 (44%)	12 (52%)		20 (55%)	
3	28 (56%)	11 (48%)		15 (42%)	
Hormone receptor status			0.24		0.66
Positive	23 (46%)	14 (61%)		24 (67%)	
Negative	27 (54%)	9 (39%)		12 (33%)	
HER2 status			0.87		0.64
Positive	10 (20%)	5 (22%)		6 (17%)	
Negative	40 (80%)	18 (78%)		30 (83%)	

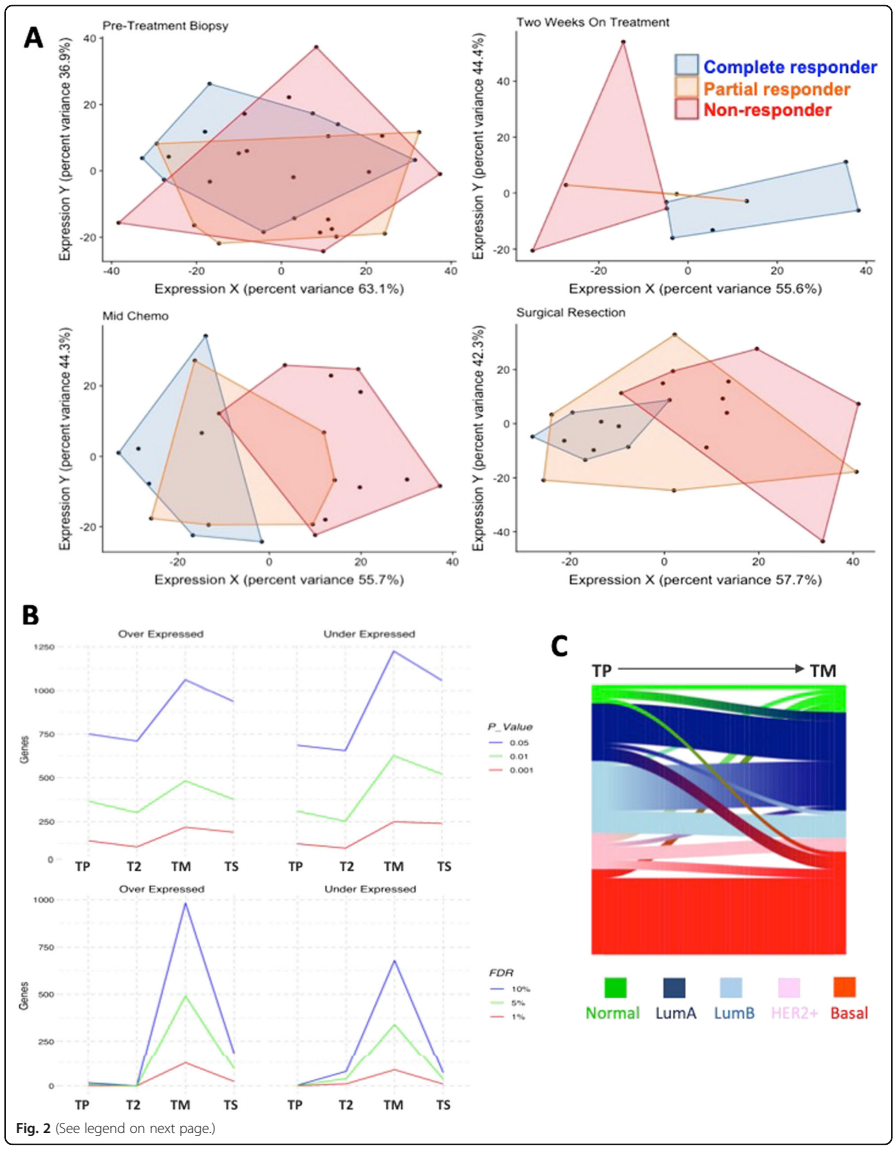


Fig. 2 (See legend on next page.)

(See figure on previous page.)
Fig. 2 Responders and non-responders are more distinct on than before treatment. **a** Supervised clustering using local Fisher discriminant analysis (LFDA) indicates that as early as 2 weeks on treatment, there is a visible separation of the response classes that were unseen in the pre-treatment samples in the NEO dataset. Red = non-responder, orange = partial responder, blue = complete responder. **b** Greater numbers of genes are under and overexpressed between responders and non-responders on treatment. The three lines represent different statistical thresholds (*$p < 0.05$, **$p < 0.01$, and ***$p < 0.001$ or FDR = 10%, FDR = 5%, and FDR = 1%, gene lists are in Additional file 4: Tables S2 and S3) in the NEO dataset. **c** Sankey diagram illustrating the proportions of tumours that change or maintain PAM50 intrinsic subtype during chemotherapy treatment. Whilst basal subtypes remain mostly stable, the composition of the cohort changes with treatment time, which may help to identify responsive or non-responsive patients. PT = pre-treatment, T-ON = on-treatment

revealed significant separation on-treatment that was not apparent in pre-treatment samples; this indicated that there are meaningful differences between the classes, as early as 2 weeks on-treatment (Fig. 2a). Complete responders and non-responsive patients were more clearly separated than partially responding patients. These results suggest that there is a potentially greater predictive value looking at on-treatment than pre-treatment biomarkers.

Responding and non-responding tumours are more different upon exposure to chemotherapy

In an attempt to quantify the molecular differences between the response groups at each time point, rank product analysis was performed at different standard p values (0.05, 0.01, and 0.001). This approach was hampered by different numbers of samples at each time point (with T2 having very few samples); however, the number of genes differentially expressed at all p values tended to be greater during rather than before treatment (Fig. 2b). Similar results were also seen using 1%, 5%, and 10% FDR (Fig. 2b). The biggest differences between the response classes were at TM (mid-chemo), which agrees with the LDFA results, which showed the least amount of overlap of the response classes at TM. Gene set enrichment analysis across the response classes at each time point also demonstrated more enriched pathways after 2 weeks of treatment (29), mid-chemo (30), and resection (29), compared to pre-treatment (18) (Additional file 2: Figure S2A). Next, we sought to examine common differentially expressed genes between responders and non-responders across the two datasets. Far more genes were commonly significantly differentially expressed (FDR = 10% between responders and non-responders on-treatment in the NEO and I-SPY 1 datasets compared with pre-treatment. In accordance with the LFDA results, more significantly differentially expressed genes (1814) were observed between on-treatment samples, with 6% (197), but only one was common between NEO and I-SPY pre-treatment (Additional file 2: Figure S2B and Additional file 4). Examination of the 468 most significantly differentially expressed genes ($p < 0.001$) between responders and non-responders in the NEO dataset at mid-chemo did not clearly distinguish between response groups or time points illustrated by the heatmap in Additional file 3: Figure S3, further demonstrating that identifying biomarkers of response to chemotherapy is very difficult.

We were also keen to evaluate whether the intrinsic subtype assigned to tumours would alter upon treatment. Looking at the NEO and I-SPY datasets, together we found that basal tumours were relatively stable with only 2/19 (11%) tumours changing. More tumours were classified as Luminal A or normal-like on-treatment, which likely reflects a reduction in the expression of proliferation genes during chemotherapy (Fig. 2c).

AAGAB is a promising potential novel on-treatment biomarker of response to chemotherapy

The mid-chemo gene list from the NEO dataset (1102 genes, unadjusted p value = 0.01) was fed to a random forest model for further feature selection and classification and regression tree (CART) model, which reported *AAGAB* as the most predictive gene for response prediction in the NEO training dataset with 100% accuracy for pCR prediction on the mid-chemo samples (Fig. 3a). Validation was conducted completely independently on publicly available sequentially sampled chemotherapy data from the I-SPY 1 Trial [10] and reported 76% accuracy using *AAGAB* at the same expression level on the scaled and centred expression data at the on-treatment time point prior to resection (T2). For comparison, the pre-treatment only sample gene lists were put through the same protocol in order to consider whether highly predictive models could be generated before chemotherapy. IGF1R was the most predictive pre-treatment marker with an accuracy of 74% and 63% in the NEO and I-SPY datasets, respectively (Table 2). AAGAB was the sixth most accurate predictor (65%, 57%); receiver operator curves show the relative specificity and sensitivity of this marker pre- and on-treatment (Fig. 3b). Gene expression levels of AAGAB were lower in responders across all time points in the NEO cohort but were most significantly different at mid-chemo. In the I-SPY dataset, *AAGAB* was significantly lower before treatment and at excision (Fig. 3c). We wondered whether *AAGAB* was lower in responders due to a reduction in proliferation, but Pearson correlation analysis

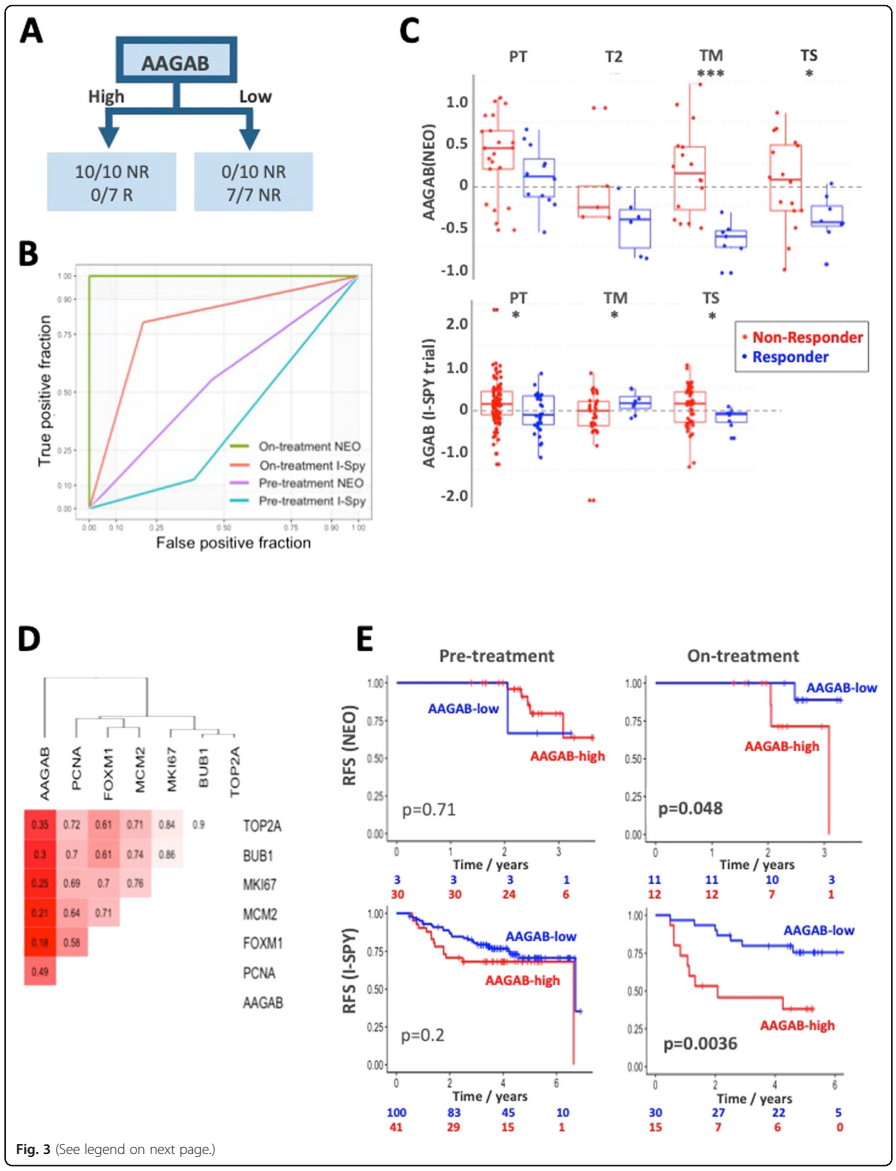


Fig. 3 (See legend on next page.)

(See figure on previous page.)
Fig. 3 AAGAB is a promising on-treatment biomarker of chemotherapy response and outcome. **a** CART analysis identified AAGAB as a possible biomarker from the Edinburgh NEO dataset and was 100% accurate at predicting pCR in the training data and 76% accurate in the I-SPY 1 validation set. **b** The ROC curves highlight the difference in on-treatment and pre-treatment accuracy and selectivity. **c** Strip charts showing the level of AAGAB in responding and non-responding patients across time points. **d** AAGAB showed no significant (Pearson) correlation with established markers of proliferation in the NEO dataset, indicating it does not seem to be a downstream proxy of their regulation. **e** Kaplan-Meier plots demonstrate that on-treatment, but not pre-treatment, levels of AAGAB were significantly associated with the outcome in both cohorts. *p* values are log-rank test

with common proliferation-associated genes (*TOP2A*, *BUB1*, *MKI67*, *MCM2*, *FOXM1*, and *PCNA*) demonstrated no significant correlation to any of these genes (Fig. 3d), suggesting that *AAGAB* is independent of proliferation. Survival analysis demonstrated that response status predicted by *AAGAB* level, at mid chemo in the NEO study and at 2 weeks in the I-SPY 1, was significantly associated with the outcome (NEO $p = 0.048$, I-SPY 1 $p = 0.0036$) (Fig. 3e). Interestingly, the level of *AAGAB* before treatment was not associated with the outcome in either cohort ($p = 0.71$ and $p = 0.2$, Fig. 3e). None of the other top 10 pre- or on-treatment markers was significantly associated with the outcome in both datasets (Table 2); only one gene (*ARF5*) was associated with the outcome in the NEO dataset ($p = 0.004$). Taken together, the single gene on-treatment biomarker AAGAB appears to outperform novel pre-treatment markers and established prognostic tests in predicting pCR and long-term outcome to chemotherapy.

Comparison of pre- and on-treatment predictions of response and outcome

We were also keen to assess whether estimations of established prognostic signatures might be different upon treatment and if on-treatment might be more accurate. All and almost all responding patients were predicted to have poor

Table 2 Comparison of pre- and on-treatment biomarkers for predicting response and outcome. Evaluation of the performance of the top 10 pre- and on-treatment genes identified for predicting pathological response in the NEO dataset

	Response accuracy		Response AUC		Outcome (log-rank)	
	NEO	I-SPY	NEO	I-SPY	NEO	I-SPY
On-treatment						
AAGAB	*100%*	78%	1.00	0.63	**0.048**	**0.0036**
ZNF165	*88%*	54%	0.91	0.57	0.26	0.70
KRTCAP3	*79%*	52%	0.85	0.56	0.81	0.49
RFC2	*79%*	40%	0.85	0.35	0.51	0.44
C20orf151	*70%*	NA	0.75	NA	0.36	NA
ARF5	*70%*	43%	0.75	0.36	**0.0038**	0.20
BSPRY	*70%*	48%	0.75	0.49	0.47	0.19
NGRN	*58%*	NA	0.66	Na	0.53	Na
CHST7	*29%*	46%	0.21	0.52	0.65	0.40
SLC18B1	*25%*	Na	0.18	NA	0.55	NA
Pre-treatment						
IGF1R	*74%*	63%	0.76	0.69	0.36	0.11
CTNNB1	*71%*	49%	0.73	0.46	0.60	0.40
SLC20A2	*71%*	56%	0.72	0.57	0.063	0.56
HMGCL	*68%*	47%	0.67	0.45	0.10	0.97
ST6GALNAC5	*68%*	52%	0.69	0.53	0.6	0.28
AAGAB	*65%*	57%	0.65	0.58	0.71	0.20
C1orf51	*62%*	NA	0.61	NA	0.12	NA
KRTCAP3	*62%*	54%	0.63	0.57	0.78	0.78
SETDB2	*50%*	49%	0.48	0.51	0.29	0.15
FADS2	*29%*	48%	0.27	0.5	0.14	0.73

NA not available, gene not represented in I-SPY dataset; *AUC* area under curve. Bold indicates significant *p*-values. Italics indicate training prediction percentages

outcomes with the estimated Mammaprint [19], PAM50 [20], or rorS [21] signatures in pre-treatment samples of the NEO cohort, whereas around half of the responding patients were predicted as good outcome using on-treatment data (Fig. 4a). Overall accuracy improved by 2–8% using on- rather than pre-treatment data; however, improvement in the predictive power of these tests was not uniform between response classes. Good outcome

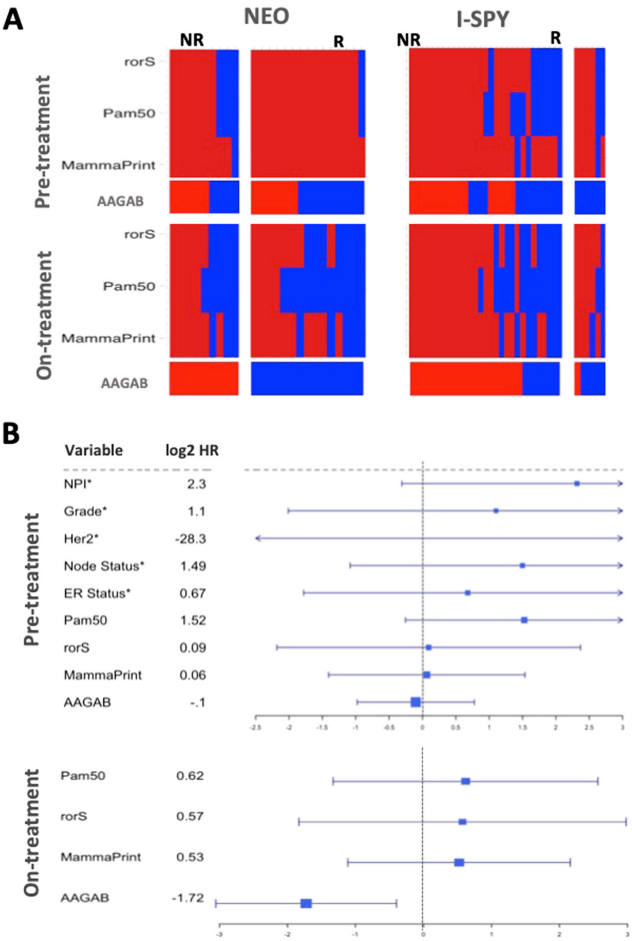


Fig. 4 On-treatment signatures more accurately predict pathological response and outcome than pre-treatment. **a** A greater proportion of patients with pathological response are predicted as responders with estimations of molecular signatures on-treatment than pre-treatment. Concordance between patients predicted as high and low risk across time is poor, but the positive predictive value of these tests increase with treatment. For PAM50 subtypes, normal-like and Luminal A are considered good prognosis and basal/Luminal B/HER2-enriched are considered poor outcome. Red = predicted poor outcome, blue = predicted good outcome. **b** Forrest plots to compare molecular signatures and AAGAB before and on-treatment combining both datasets, except where indicated* due to individual sample data unavailable for I-SPY 1 patients

predictions for responders to neoadjuvant chemotherapy saw an aggregate increase in predictive power from 11 to 44.4%, whilst poor outcome predictions for non-responders saw a moderate decrease in accuracy, 75 to 63%. None of the gene expression signatures either pre- or on-treatment or established prognostic markers (NPI, Grade, Her2 status) was significantly associated with the outcome in contrast to the remarkable performance of on-treatment measurement of AAGAB (Fig. 4b).

Discussion

Determining molecular differences between tumours to select the most effective treatment is the defining feature of precision oncology. Accurately predicting which patients will respond to treatment before exposure relies on a highly specific target. In breast cancer, ER status is a good indicator of response to endocrine treatment, but resistance, both primary and acquired, is common. Chemotherapy is an unselective treatment, relying on cancer cells growing faster than normal cells. The results presented here, along with others [7, 8], suggest on-treatment biomarkers have improved value in predicting whether tumours respond to treatment and are associated with the outcome. Changes in gene expression in sequential patient-matched were fairly consistent in response to chemotherapy across two independent datasets, regardless of the response status. Identifying molecular markers between responding and non-responding tumours was much more challenging. We previously demonstrated that lobular and ductal breast cancers respond to endocrine treatment in the same way, despite clear histological and molecular distinctions that are apparent and maintained on-treatment [22], demonstrating that pre-treatment variations do not necessarily lead to differences in response. The results of this study are somewhat exploratory, rather than definitive, but further illustrate the considerable potential value of on-treatment sampling.

There are no universally agreed-upon markers predictive of response to chemotherapy, and the few that have been investigated in the neoadjuvant setting typically centre around established markers including ER, P53 HER2, and Ki-67 [23]; thus, the introduction of new novel biomarkers can expand the currently available clinical options for physicians. A study published over a decade ago stated that the differences in gene expression between responders and non-responders to neoadjuvant chemotherapy must be rather subtle [12]. The results presented here confirm this statement; however, our results suggest that on-treatment biomarkers may provide important information for predicting response.

As cancer is inherently a proliferative disease, measuring the change in markers of proliferation on-treatment is logical and genes like ki-67 have been demonstrated previously to be potentially a new clinical tool for disease prognosis and prediction [24, 25]. It is therefore all the more interesting that the potentially novel biomarker identified in this study, AAGAB is not tightly correlated with known markers of proliferation. AAGAB has primarily been studied for its role in punctate palmoplantar keratoderma [26] and the role of adaptin in the clathrin-independent endocytosis of epidermal growth factors. The level of AAGAB was found to be prognostic of response ($p < 0.001$) in renal cancers (favourably) and in thyroid cancers (unfavourably) from the TCGA study, and expression is elevated in breast cancer, relative to the normal breast ($p < 0.001$). However, the exact role of AAGAB in breast cancer is currently unclear and potentially warrants further investigation. Clearly, further validation of the role of AAGAB in breast cancer is warranted and will be performed as new neoadjuvant chemotherapy datasets become available. This study supports the use and identification of genes or markers from on-treatment biopsies as a tool for improving patient response classification. We propose that the use of on-treatment samples offers valuable insight into the dynamic changes correlated with response, and submit our findings as support for continued neoadjuvant sampling, and novel biomarker generation.

Conclusion

We have identified AAGAB as a novel on-treatment biomarker for accurate prediction of pCR and outcome in patients treated with neoadjuvant chemotherapy. A semi-supervised analysis and evaluation of estimations of established molecular signatures also highlight the potential value of on-treatment biomarkers. Combining on-treatment biomarkers with known clinical prognostic factors could further improve the accuracy of response predictions and deserve further study. On-treatment expression changes in the neoadjuvant setting may offer greater possibilities for the identification and creation of more future novel biomarkers.

Additional files

Additional file 1: Figure S1. A, Pairwise significant analysis of microarrays (FDR = 10%) demonstrating that whilst only a relatively small proportion of genes are significantly up- or downregulated in response to chemotherapy in both datasets, overall changes in patient-matched sequential samples response to treatment are highly consistent. Red = upregulated, blue = downregulated on- relative to pre-treatment. Gene lists are in Additional file 4: Table S1. B, Unsupervised principal component analysis cannot distinguish responding from non-responding breast tumours receiving chemotherapy, before or on-treatment. (JPG 126 kb)

Additional file 2: Figure S2. A, Gene set enrichment analysis (GSEA) results showing greater numbers of enriched pathways between responders and non-responders on-treatment compared to pre-treatment in the NEO dataset. B, Venn diagrams indicating that there were many more overlapping significantly differentially expressed genes

between responders and non-responders across the two studies on-treatment compared to pre-treatment. Gene lists are for FDR = 10% (see Additional file 4: Table S3). (JPG 72 kb)

Additional file 3: Figure S3. Heatmap of the 468 most significantly differentially expressed genes ($p < 0.001$) between responders and non-responders in the NEO dataset at mid-chemo, demonstrating rather poor separation between the response groups and time points. Gene list is in Additional file 4: Table S2. (JPG 91 kb)

Additional file 4: Table S1. Gene lists of pairwise analysis of pre- and on-treatment sequential patient-matched samples from NEO and I-SPY datasets using Significance analysis of microarrays with FDR = 10%. **Table S2.** Gene lists distinguishing between responders and non-responders at different time points across the NEO and ISPY datasets using rank product analysis with $p < 0.05$, $p < 0.01$, and $p < 0.001$. Pre-treatment (TP), 2 weeks (T2), mid-chemo (TM), and surgery (TS). **Table S3.** Gene lists distinguishing between responders and non-responders at different time points across the NEO and I-SPY datasets using significance analysis of microarrays with FDR = 10%, FDR = 5%, and FDR = 1%. Pre-treatment (TP), 2 weeks (T2), mid-chemo (TM), and surgery (TS). (XLSX 1184 kb)

Acknowledgements
We are grateful for the help and support of the Wellcome Trust Clinical Research Facility under the direction of Lee Murphy.

Authors' contributions
OO conceived the study. OO recruited the patients and collected all samples and relevant clinical information. AKT and CM-P co-ordinated the samples and generated the NEO transcriptome dataset. RJB performed the data analysis. DAC, AHS, and OO interpreted the data. RJB and AHS drafted the manuscript. All authors read and approved the final manuscript.

Authors' information
Not applicable

Funding
Cancer Research UK is thanked for providing funding for a PhD stipend and a development fund to the Edinburgh CRUK Centre. AHS is grateful for the funding from Breast Cancer Now.

Availability of data and materials
Gene expression data associated from the study has been made publicly available at NCBI GEO under the accession number GSE122630.

Ethics approval and consent to participate
The protocol for the "NEO study: a study of factors predicting response to neoadjuvant chemotherapy in breast cancer" was approved by the South East Scotland Research Ethics Committee 01 with REC reference: 13/SS/0236. All patients signed an informed consent at least 24 h after receiving the patient information leaflet and after having the chance to discuss thoroughly their participation to this study either with treating clinician or research nurses.

Consent for publication
Not applicable

Competing interests
All authors declare that they have no competing interests.

Author details
[1]University of Edinburgh Cancer Research UK Centre, MRC Institute of Genetics and Molecular Medicine, Edinburgh, UK. [2]Edinburgh Cancer Centre, Western General Hospital, Edinburgh, UK.

Received: 5 December 2018 Accepted: 5 June 2019
Published online: 14 June 2019

References

1. Schott AF, Hayes DF. Defining the benefits of neoadjuvant chemotherapy for breast cancer. J Clin Oncol. 2012;30:1747–9.
2. Esserman LJ, Berry DA, DeMichele A, Carey L, Davis SE, Buxton M, Hudis C, Gray JW, Perou C, Yau C, Livasy C, Krontiras H, Montgomery L, Tripathy D, Lehman C, Liu MC, Olopade OI, Rugo HS, Carpenter JT, Dressler L, Chhieng D, Singh B, Mies C, Rabban J, Chen Y-Y, Giri D, van 't Veer L, Hylton N. Pathologic complete response predicts recurrence-free survival more effectively by cancer subset: results from the I-SPY 1 TRIAL–CALGB 150007/150012, ACRIN 6657. J Clin Oncol. 2012;30:3242–9.
3. Untch M, Konecny GE, Paepke S, von Minckwitz G. Current and future role of neoadjuvant therapy for breast cancer. Breast. 2014;23:526–37.
4. Early Breast Cancer Trialists' Collaborative Group (EBCTCG). Long-term outcomes for neoadjuvant versus adjuvant chemotherapy in early breast cancer: meta-analysis of individual patient data from ten randomised trials. Lancet Oncol. 2018;19:27–39.
5. Sims AH, Bartlett JM. Approaches towards expression profiling the response to treatment. Breast Cancer Res. 2008;10:115.
6. Macaskill EJ, Dixon JM. Neoadjuvant endocrine therapy. In: Kuerer HM, editor. Breast surgical oncology. New York: McGraw Hill Medical; 2010.
7. Turnbull AK, Arthur LM, Renshaw L, Larionov AA, Kay C, Dunbier AK, Thomas JS, Dowsett M, Sims AH, Dixon JM. Accurate prediction and validation of response to endocrine therapy in breast cancer. J Clin Oncol. 2015;33:2270–8.
8. Ellis MJ, Suman VJ, Hoog J, Goncalves R, Sanati S, Creighton CJ, DeSchryver K, Crouch E, Brink A, Watson M, Luo J, Tao Y, Barnes M, Dowsett M, Budd GT, Winer E, Silverman P, Esserman L, Carey L, Ma CX, Unzeitig G, Pluard T, Whitworth P, Babiera G, Guenther JM, Dayao Z, Ota D, Leitch M, Olson JA, Allred DC, et al. Ki67 proliferation index as a tool for chemotherapy decisions during and after neoadjuvant aromatase inhibitor treatment of breast cancer: results from the American College of Surgeons Oncology Group Z1031 Trial (Alliance). J Clin Oncol. 2017;35:1061–9.
9. Sorlie T, Perou CM, Fan C, Geisler S, Aas T, Nobel A, Anker G, Akslen LA, Botstein D, Borresen-Dale AL, Lonning PE. Gene expression profiles do not consistently predict the clinical treatment response in locally advanced breast cancer. Mol Cancer Ther. 2006;5:2914–8.
10. Hatzis C, Pusztai L, Valero V, Booser DJ, Esserman L, Lluch A, Vidaurre T, Holmes F, Souchon E, Wang H, Martin M, Cotrina J, Gomez H, Hubbard R, Chacón JI, Ferrer-Lozano J, Dyer R, Buxton M, Gong Y, Wu Y, Ibrahim N, Andreopoulou E, Ueno NT, Hunt K, Yang W, Nazario A, DeMichele A, O'Shaughnessy J, Hortobagyi GN, Symmans WF. A genomic predictor of response and survival following taxane-anthracycline chemotherapy for invasive breast cancer. JAMA. 2011;305:1873–81.
11. Magbanua MJM, Wolf DM, Yau C, Davis SE, Crothers J, Au A, Haqq CM, Livasy C, Rugo HS, Esserman L, Park JW, van 't Veer LJ. Serial expression analysis of breast tumors during neoadjuvant chemotherapy reveals changes in cell cycle and immune pathways associated with recurrence and response. Breast Cancer Res. 2015;17:73.
12. Hannemann J, Oosterkamp HM, Bosch CAJ, Velds A, Wessels LFA, Loo C, Rutgers EJ, Rodenhuis S, van de Vijver MJ. Changes in gene expression associated with response to neoadjuvant chemotherapy in breast cancer. J Clin Oncol. 2005;23:3331–42.
13. Gonzalez-Angulo AM, Iwamoto T, Liu S, Chen H, Do KA, Hortobagyi GN, Mills GB, Meric-Bernstam F, Symmans WF, Pusztai L. Gene expression, molecular class changes, and pathway analysis after neoadjuvant systemic therapy for breast cancer. Clin Cancer Res. 2012;18:1109–19.
14. Abraham A, Pedregosa F, Eickenberg M, Gervais P, Muller A, Kossaifi J, Gramfort A, Thirion B, Varoquaux G. Machine learning for neuroimaging with scikit-learn. Front Neuroinform. 2014;8:14.
15. Smyth GK, Michaud J, Scott HS. Use of within-array replicate spots for assessing differential expression in microarray experiments. Bioinformatics. 2005;21:2067–75.
16. Tang Y, Li W. lfda: an R package for local Fisher discriminant analysis and visualization; 2016.
17. Tusher VG, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. Proc Natl Acad Sci. 2001;98:5116–21.
18. Gendoo DM, Ratanasirigulchai N, Schröder MS, Paré L, Parker JS, Prat A, Haibe-Kains B. Genefu: an R/Bioconductor package for computation of gene expression-based signatures in breast cancer. Bioinformatics. 2016;32(7):1097–9.
19. van 't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AA, Mao M, Peterse HL, van der Kooy K, Marton MJ, Witteveen AT, Schreiber GJ, Kerkhoven RM, Roberts C, Linsley PS, Bernards R, Friend SH. Gene expression profiling predicts clinical outcome of breast cancer. Nature. 2002;415:530.

20. Parker JS, Mullins M, Cheang MC, Leung S, Voduc D, Vickery T, Davies S, Fauron C, He X, Hu Z, Quackenbush JF, Stijleman IJ, Palazzo J, Marron JS, Nobel AB, Mardis E, TO N, Ellis MJ, Perou CM, Bernard PS. Supervised risk predictor of breast cancer based on intrinsic subtypes. J Clin Oncol. 2009;27:1160.
21. Wallden B, Storhoff J, Nielsen T, Dowidar N, Schaper C, Ferree S, Liu S, Leung S, Geiss G, Snider J, Vickery T, Davies SR, Mardis ER, Gnant M, Sestak I, Ellis MJ, Perou CM, Bernard PS, Parker JS. Development and verification of the PAM50-based Prosigna breast cancer gene signature assay. BMC Med Genet. 2015;8:54.
22. Arthur LM, Turnbull AK, Webber VL, Larionov AA, Renshaw L, Kay C, Thomas JS, Dixon JM, Sims AH. Molecular changes in lobular breast cancers in response to endocrine therapy. Cancer Res. 2014;74:5371–6.
23. Tewari M, Krishnamurthy A, Shukla HS. Predictive markers of response to neoadjuvant chemotherapy in breast cancer. Surg Oncol. 2008;17:301–11.
24. Colozza M, Azambuja E, Cardoso F, Sotiriou C, Larsimont D, Piccart MJ. Proliferative markers as prognostic and predictive tools in early breast cancer: where are we now? Ann Oncol Off J Eur Soc Med Oncol. 2005;16:1723–39.
25. Urruticoechea A, Smith IE, Dowsett M. Proliferation marker Ki-67 in early breast cancer. J Clin Oncol. 2005;23:7212–20.
26. Pohler E, Mamai O, Hirst J, Zamiri M, Horn H, Nomura T, Irvine AD, Moran B, Wilson NJ, Smith FJD, Goh CSM, Sandilands A, Cole C, Barton GJ, Evans AT, Shimizu H, Akiyama M, Suehiro M, Konohana I, Shboul M, Teissier S, Boussofara L, Denguezli M, Saad A, Gribaa M, Dopping-Hepenstal PJ, McGrath JA, Brown SJ, Goudie DR, Reversade B, et al. Haploinsufficiency for AAGAB causes clinically heterogeneous forms of punctate palmoplantar keratoderma. Nat Genet. 2012;44:1272–6.

Publisher's Note

References

1. Hanahan D, Weinberg RA. The Hallmarks of Cancer. Cell [Internet]. 2000 Jan 7 [cited 2018 Mar 16];100(1):57–70. Available from: http://www.sciencedirect.com/science/article/pii/S0092867400816839

2. Cheung-Ong K, Giaever G, Nislow C. DNA-Damaging Agents in Cancer Chemotherapy: Serendipity and Chemical Biology. Chemistry & Biology [Internet]. 2013 May [cited 2018 Mar 16];20(5):648–59. Available from: http://linkinghub.elsevier.com/retrieve/pii/S1074552113001312

3. Raguz S, Yagüe E. Resistance to chemotherapy: New treatments and novel insights into an old problem. British Journal of Cancer [Internet]. 2008 Aug [cited 2018 Mar 16];99(3):387–91. Available from: https://www.nature.com/articles/6604510

4. Hayes EL, Lewis-Wambi JS. Mechanisms of endocrine resistance in breast cancer: an overview of the proposed roles of noncoding RNA. Breast cancer research : BCR [Internet]. 2015;17(1):542. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4362832%7B/&%7Dtool=pmcentrez%7B/&%7Drendertype=abstract

5. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: a cancer journal for clinicians. 2018;68(6):394–424.

6. Hayes EL, Lewis-Wambi JS. Mechanisms of endocrine resistance in breast cancer: An overview of the proposed roles of noncoding rna. Breast Cancer Research [Internet]. 2015 Mar;17(1). Available from: http://dx.doi.org/10.1186/s13058-015-0542-y

7. Torre LA, Bray F, Siegel RL, Ferlay J, Lortet-Tieulent J, Jemal A. Global cancer statistics, 2012. CA: A Cancer Journal for Clinicians [Internet]. 2015 Mar;65(2):87–108. Available from: http://www.ncbi.nlm.nih.gov/pubmed/25651787%20http://doi.wiley.com/10.3322/caac.21262

8. Breast cancer statistics. http://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/breast-cancer#heading-Two;

9. Of UTOED, Group BC. First results on mortality reduction in the uk trial of early detection of breast cancer. The Lancet. 1988;332(8608):411–6.

10. Cancer stat facts: Female breast cancer. https://seer.cancer.gov/statfacts/html/breast.html;

11. Ferlay J SI DR Ervik M. Cancer incidence and mortality worldwide. http://globocan.iarc.fr;

12. Torre LA, Siegel RL, Ward EM, Jemal A. Global cancer incidence and mortality rates and trends–an update. Cancer Epidemiology Biomarkers & Prevention [Internet]. 2015 Dec;25(1):16–27. Available from: http://dx.doi.org/10.1158/1055-9965.EPI-15-

0578

13. Polyak K. Heterogeneity in breast cancer. Journal of Clinical Investigation [Internet]. 2011 Oct;121(10):3786–8. Available from: http://dx.doi.org/10.1172/JCI60534

14. Stephens PJ, Tarpey PS, Davies H, Van Loo P, Greenman C, Wedge DC, et al. The landscape of cancer genes and mutational processes in breast cancer. Nature [Internet]. 2012 May;486(7403):400–4. Available from: http://dx.doi.org/10.1038/nature11017

15. Breast cancer screening. http://www.cancerresearchuk.org/about-cancer/breast-cancer/screening/breast-screening;

16. Saslow D, Hannan J, Osuch J, Alciati MH, Baines C, Barton M, et al. Clinical breast examination: practical recommendations for optimizing performance and reporting. CA: a cancer journal for clinicians [Internet]. 54(6):327–44. Available from: http://www.ncbi.nlm.nih.gov/pubmed/15537576

17. Dackus G, Hoeve ND ter, Opdam M, Vreuls W, Koop EA, Willems SM, et al. Long-term outcome of breast cancer patients diagnosed ?40 years according to breast cancer subtype in the absence of adjuvant systemic therapy: The paradigm initiative. Journal of Clinical Oncology [Internet]. 2017;35(15_suppl):535–5. Available from: http://ascopubs.org/doi/abs/10.1200/JCO.2017.35.15_suppl

18. Malhotra GK, Zhao X, Band H, Band V. Histological, molecular and functional subtypes of breast cancers. Cancer biology & therapy. 2010;10(10):955–60.

19. What is breast cancer? https://www.cancer.org/cancer/breast-cancer/about/what-is-breast-cancer.html;

20. Types of breast cancer. http://www.breastcancer.org/symptoms/types;

21. Connolly J LV Kempson R. Recommendations for the reporting of breast carcinoma. Association of Directors of Anatomic and Surgical Pathology. 2004;

22. Li C, Uribe D, Daling J. Clinical characteristics of different histologic types of breast cancer. British journal of cancer. 2005;93(9):1046.

23. Vallejos CS, Gómez HL, Cruz WR, Pinto JA, Dyer RR, Velarde R, et al. Breast Cancer Classification According to Immunohistochemistry Markers: Subtypes and Association With Clinicopathologic Variables in a Peruvian Hospital Database. Clinical Breast Cancer [Internet]. 2010 Aug;10(4):294–300. Available from: http://www.ncbi.nlm.nih.gov/pubmed/20705562%20http://linkinghub.elsevier.com/retrieve/pii/S1526820911700416

24. Dai X, Li T, Bai Z, Yang Y, Liu X, Zhan J, et al. Breast cancer intrinsic subtype classification, clinical use and future trends. American journal of cancer research [Internet]. 2015;5(10):2929–43. Available from: http://www.ncbi.nlm.nih.gov/pubmed/26693050%20http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4656721

25. Sorlie T, Perou CM, Tibshirani R, Aas T, Geisler S, Johnsen H, et al. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. Proceedings of the National Academy of Sciences [Internet]. 2001 Sep;98(19):10869–74. Available from: http://www.ncbi.nlm.nih.gov/pubmed/11553815%20http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC58566%20http://www.pnas.org/cgi/doi/10.1073/pnas.191367098

26. Brenton JD, Carey LA, Ahmed AA, Caldas C. Molecular Classification and Molecular Forecasting of Breast Cancer: Ready for Clinical Application? Journal of Clinical Oncology [Internet]. 2005 Oct;23(29):7350–60. Available from: http://www.ncbi.nlm.nih.gov/pubmed/16145060%20http://ascopubs.org/doi/10.1200/JCO.2005.03.3845

27. Perou CM, Sørlie T, Eisen MB, Rijn M van de, Jeffrey SS, Rees CA, et al. Molecular portraits of human breast tumours. Nature [Internet]. 2000 Aug;406(6797):747–52. Available from: http://www.ncbi.nlm.nih.gov/pubmed/10963602%20http://www.nature.com/doifinder/10.1038/35021093

28. BREAST cancer. http://www.pathophys.org/breast-cancer/#Pathogenesis_hormone_sensitive;

29. Teschendorff AE, Miremadi A, Pinder SE, Ellis IO, Caldas C. An immune response gene expression module identifies a good prognosis subtype in estrogen receptor negative breast cancer. Genome Biology [Internet]. 2007;8(8):R157. Available from: http://genomebiology.biomedcentral.com/articles/10.1186/gb-2007-8-8-r157

30. Lehmann BD, Jovanović B, Chen X, Estrada MV, Johnson KN, Shyr Y, et al. Refinement of Triple-Negative Breast Cancer Molecular Subtypes: Implications for Neoadjuvant Chemotherapy Selection. Sapino A, editor. PLOS ONE [Internet]. 2016 Jun;11(6):e0157368. Available from: http://dx.plos.org/10.1371/journal.pone.0157368

31. Lehmann BD, Bauer JA, Chen X, Sanders ME, Chakravarthy AB, Shyr Y, et al. Identification of human triple-negative breast cancer subtypes and preclinical models for selection of targeted therapies. Journal of Clinical Investigation [Internet]. 2011 Jul;121(7):2750–67. Available from: http://www.jci.org/articles/view/45014

32. Mehrgou A, Akouchekian M. The importance of brca1 and brca2 genes mutations in breast cancer development. Medical journal of the Islamic Republic of Iran. 2016;30:369.

33. Dhankhar R, Vyas SP, Jain AK, Arora S, Rath G, Goyal AK. Advances in Novel Drug Delivery Strategies for Breast Cancer Therapy. Artificial Cells, Blood Substitutes, and Biotechnology [Internet]. 2010 Oct;38(5):230–49. Available from: http://www.ncbi.nlm.nih.gov/pubmed/20677900%20http://www.tandfonline.com/doi/full/10.3109/10731199.2010.494578

34. Matsen CB, Neumayer LA. Breast Cancer. JAMA Surgery [Internet]. 2013 Oct;148(10):971. Available from: http://www.ncbi.nlm.nih.gov/pubmed/23986370%20http://archsurg.jamanetwork.com/article.aspx?doi=10.1001/jamasurg.2013.3393

35. Board PATE. Breast cancer treatment (pdq). 2017;

36. Miller E, Lee HJ, Lulla A, Hernandez L, Gokare P, Lim B. Current treatment of early breast cancer: Adjuvant and neoadjuvant therapy. F1000Research. 2014;3.

37. Nounou MI, ElAmrawy F, Ahmed N, Abdelraouf K, Goda S, Syed-Sha-Qhattal H. Breast Cancer: Conventional Diagnosis and Treatment Modalities and Recent Patents and Technologies. Breast cancer : basic and clinical research [Internet]. 2015;9(Suppl 2):17–34. Available from: http://www.ncbi.nlm.nih.gov/pubmed/26462242%20http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4589089

38. Goldhirsch A, Winer EP, Coates AS, Gelber RD, Piccart-Gebhart M, Thürlimann B, et al. Personalizing the treatment of women with early breast cancer: highlights of the St Gallen International Expert Consensus on the Primary Therapy of Early Breast Cancer 2013. Annals of Oncology [Internet]. 2013 Sep;24(9):2206–23. Available from: http://www.ncbi.nlm.nih.gov/pubmed/23917950%20http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3755334%20https://academic.oup.com/annonc/article-lookup/doi/10.1093/annonc/mdt303

39. Akram M, Siddiqui S. Breast cancer management: Past, present and evolving.

Indian Journal of Cancer [Internet]. 2012;49(3):277. Available from: http://www.ncbi.nlm.nih.gov/pubmed/23238144%20http://www.indianjcancer.com/text.asp?2012/49/3/277/104486

40. Shao N, Wang S, Yao C, Xu X, Zhang Y, Zhang Y, et al. Sequential versus concurrent anthracyclines and taxanes as adjuvant chemotherapy of early breast cancer: A meta-analysis of phase III randomized control trials. The Breast [Internet]. 2012 Jun;21(3):389–93. Available from: http://www.ncbi.nlm.nih.gov/pubmed/22542064%20http://linkinghub.elsevier.com/retrieve/pii/S0960977612000707

41. Early Breast Cancer Trialists' Collaborative Group (EBCTCG), Peto R, Davies C, Godwin J, Gray R, Pan HC, et al. Comparisons between different polychemotherapy regimens for early breast cancer: meta-analyses of long-term outcome among 100?000 women in 123 randomised trials. The Lancet [Internet]. 2012 Feb;379(9814):432–44. Available from: http://www.ncbi.nlm.nih.gov/pubmed/22152853%20http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3273723%20http://linkinghub.elsevier.com/retrieve/pii/S0140673611616255

42. Valero V. Docetaxel and cyclophosphamide in patients with advanced solid tumors. Oncology (Williston Park, NY) [Internet]. 1997 Jun;11(6 Suppl 6):21–3. Available from: http://www.ncbi.nlm.nih.gov/pubmed/9213323

43. Hirsimäki P, Aaltonen A, Mäntylä E. Toxicity of antiestrogens. The breast journal [Internet]. 8(2):92–6. Available from: http://www.ncbi.nlm.nih.gov/pubmed/11896754

44. Puhalla S, Brufsky A, Davidson N. Adjuvant endocrine therapy for premenopausal women with breast cancer. The Breast [Internet]. 2009 Oct;18:S122–30. Available from: http://www.ncbi.nlm.nih.gov/pubmed/19914530%20http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3901991%20http://linkinghub.elsevier.com/retrieve/pii/S0960977609702863

45. Dowsett M, Stein RC, Coombes RC. Aromatization inhibition alone or in combination with GnRH agonists for the treatment of premenopausal breast cancer patients. The Journal of steroid biochemistry and molecular biology [Internet]. 1992 Sep;43(1-3):155–9. Available from: http://www.ncbi.nlm.nih.gov/pubmed/1388047

46. Davies C, Pan H, Godwin J, Gray R, Arriagada R, Raina V, et al. Long-term effects of continuing adjuvant tamoxifen to 10 years versus stopping at 5 years after diagnosis of oestrogen receptor-positive breast cancer: ATLAS, a randomised trial. The Lancet [Internet]. 2013 Mar;381(9869):805–16. Available from: http://www.ncbi.nlm.nih.gov/pubmed/23219286%20http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3596060%20http://linkinghub.elsevier.com/retrieve/pii/S0140673612619631

47. Callahan R, Hurvitz S. Human epidermal growth factor receptor-2-positive breast cancer: Current management of early, advanced, and recurrent disease. Current opinion in obstetrics & gynecology [Internet]. 2011 Feb;23(1):37–43. Available from: http://www.ncbi.nlm.nih.gov/pubmed/21500375%20http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4307801

48. Tsang RY, Finn RS. Beyond trastuzumab: novel therapeutic strategies in HER2-positive metastatic breast cancer. British Journal of Cancer [Internet]. 2012 Jan;106(1):6–13. Available from: http://www.ncbi.nlm.nih.gov/pubmed/22215104%20http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3251862%20http:

//www.nature.com/doifinder/10.1038/bjc.2011.516

49. O'Sullivan CC, Bradbury I, Campbell C, Spielmann M, Perez EA, Joensuu H, et al. Efficacy of Adjuvant Trastuzumab for Patients With Human Epidermal Growth Factor Receptor 2?Positive Early Breast Cancer and Tumors ? 2 cm: A Meta-Analysis of the Randomized Trastuzumab Trials. Journal of Clinical Oncology [Internet]. 2015 Aug;33(24):2600–8. Available from: http://www.ncbi.nlm.nih.gov/pubmed/26101239%20http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4534523%20http://ascopubs.org/doi/10.1200/JCO.2015.60.8620

50. Ewer SM, Ewer MS. Cardiotoxicity profile of trastuzumab. Drug safety [Internet]. 2008;31(6):459–67. Available from: http://www.ncbi.nlm.nih.gov/pubmed/18484781

51. Turner N, Biganzoli L, Di Leo A. Continued value of adjuvant anthracyclines as treatment for early breast cancer. The Lancet Oncology [Internet]. 2015 Jul;16(7):e362–9. Available from: http://www.ncbi.nlm.nih.gov/pubmed/26149888%20http://linkinghub.elsevier.com/retrieve/pii/S1470204515000790

52. Vinayak S, Carlson RW. mTOR inhibitors in the treatment of breast cancer. Oncology (Williston Park, NY) [Internet]. 2013 Jan;27(1):38–44, 46, 48 passim. Available from: http://www.ncbi.nlm.nih.gov/pubmed/23461041

53. Mayer EL. Targeting Breast Cancer with CDK Inhibitors. Current Oncology Reports [Internet]. 2015 May;17(5):20. Available from: http://www.ncbi.nlm.nih.gov/pubmed/25716100%20http://link.springer.com/10.1007/s11912-015-0443-3

54. Senkus E, Kyriakides S, Ohno S, Penault-Llorca F, Poortmans P, Rutgers E, et al. Primary breast cancer: ESMO clinical practice guidelines for diagnosis, treatment and follow-up. Annals of oncology. 2015;26(suppl_5):v8–v30.

55. Haybittle JL, Blamey RW, Elston CW, Johnson J, Doyle PJ, Campbell FC, et al. A prognostic index in primary breast cancer. British journal of cancer [Internet]. 1982 Mar;45(3):361–6. Available from: http://www.ncbi.nlm.nih.gov/pubmed/7073932%20http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2010939

56. Wishart GC, Azzato EM, Greenberg DC, Rashbass J, Kearins O, Lawrence G, et al. PREDICT: a new UK prognostic model that predicts survival following surgery for invasive breast cancer. Breast Cancer Research [Internet]. 2010 Feb;12(1):R1. Available from: http://breast-cancer-research.biomedcentral.com/articles/10.1186/bcr2464

57. Glas NA de, Water W van de, Engelhardt EG, Bastiaannet E, Craen AJM de, Kroep JR, et al. Validity of Adjuvant! Online program in older patients with breast cancer: a population-based study. The Lancet Oncology [Internet]. 2014 Jun;15(7):722–9. Available from: http://www.ncbi.nlm.nih.gov/pubmed/24836274

58. Sorlie T, Tibshirani R, Parker J, Hastie T, Marron J, Nobel A, et al. Repeated observation of breast tumor subtypes in independent gene expression data sets. Proceedings of the National Academy of Sciences of the United States of America. 2003;100(14):8418–23.

59. Harris L, Fritsche H, Mennel R, Norton L, Ravdin P, Taube S, et al. American society of clinical oncology 2007 update of recommendations for the use of tumor markers in breast cancer. Journal of clinical oncology. 2007;25(33):5287–312.

60. Perou CM, Sorlie T, Eisen MB, Van De Rijn M, others. Molecular portraits of human breast tumours. nature. 2000;406(6797):747.

61. Sørlie T, Perou CM, Tibshirani R, Aas T, Geisler S, Johnsen H, et al. Gene expression

patterns of breast carcinomas distinguish tumor subclasses with clinical implications. Proceedings of the National Academy of Sciences. 2001;98(19):10869–74.

62. Parker JS, Mullins M, Cheang MC, Leung S, Voduc D, Vickery T, et al. Supervised risk predictor of breast cancer based on intrinsic subtypes. Journal of clinical oncology. 2009;27(8):1160–7.

63. Sorlie T, Tibshirani R, Parker J, Hastie T, Marron JS, Nobel A, et al. Repeated observation of breast tumor subtypes in independent gene expression data sets. Proceedings of the National Academy of Sciences [Internet]. 2003 Jul;100(14):8418–23. Available from: http://www.ncbi.nlm.nih.gov/pubmed/12829800%20http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC166244%20http://www.pnas.org/cgi/doi/10.1073/pnas.0932692100

64. Hon JDC, Singh B, Sahin A, Du G, Wang J, Wang VY, et al. Breast cancer molecular subtypes: from TNBC to QNBC. American journal of cancer research [Internet]. 2016;6(9):1864–72. Available from: http://www.ncbi.nlm.nih.gov/pubmed/27725895%20http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5043099

65. Clare SE, Shaw PL. ?Big Data? for breast cancer: where to look and what you will find. npj Breast Cancer [Internet]. 2016 Dec;2(1):16031. Available from: http://www.nature.com/articles/npjbcancer201631

66. Pereira B, Chin S-F, Rueda OM, Vollan H-KM, Provenzano E, Bardwell HA, et al. The somatic mutation profiles of 2,433 breast cancers refines their genomic and transcriptomic landscapes. Nature Communications [Internet]. 2016 May;7:11479. Available from: http://www.ncbi.nlm.nih.gov/pubmed/27161491%20http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4866047%20http://www.nature.com/doifinder/10.1038/ncomms11479

67. Ali HR, Rueda OM, Chin S-F, Curtis C, Dunning MJ, Aparicio SA, et al. Genome-driven integrated classification of breast cancer validated in over 7,500 samples. Genome Biology [Internet]. 2014 Aug;15(8):431. Available from: http://genomebiology.biomedcentral.com/articles/10.1186/s13059-014-0431-1

68. Koboldt DC, Fulton RS, McLellan MD, Schmidt H, Kalicki-Veizer J, McMichael JF, et al. Comprehensive molecular portraits of human breast tumours. Nature [Internet]. 2012 Sep;490(7418):61–70. Available from: http://www.ncbi.nlm.nih.gov/pubmed/23000897%20http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3465532%20http://www.nature.com/doifinder/10.1038/nature11412

69. Galperin MY, Fernández-Suárez XM, Rigden DJ. The 24th annual Nucleic Acids Research database issue: a look back and upcoming changes. Nucleic acids research [Internet]. 2017;45(9):5627. Available from: http://www.ncbi.nlm.nih.gov/pubmed/28100696%20http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5435940

70. Rigden DJ, Fernández-Suárez XM, Galperin MY. The 2016 database issue of Nucleic Acids Research and an updated molecular biology database collection. Nucleic acids research [Internet]. 2016 Jan;44(D1):D1–6. Available from: http://www.ncbi.nlm.nih.gov/pubmed/26740669%20http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4702933

71. Galperin MY, Rigden DJ, Fernández-Suárez XM. The 2015 Nucleic Acids Research Database Issue and Molecular Biology Database Collection. Nucleic Acids Research [Internet]. 2015 Jan;43(D1):D1–5. Available from: http://www.ncbi.nlm.nih.gov/pubmed/

25593347%20http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4383995%20http://academic.oup.com/nar/article/43/D1/D1/2439464/The-2015-Nucleic-Acids-Research-Database-Issue-and

72. Fernández-Suárez XM, Rigden DJ, Galperin MY. The 2014 Nucleic Acids Research Database Issue and an updated NAR online Molecular Biology Database Collection. Nucleic Acids Research [Internet]. 2014 Jan;42(D1):D1–6. Available from: http://www.ncbi.nlm.nih.gov/pubmed/24316579%20http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3965027%20https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkt1282

73. Gendoo DM, Zon M, Sandhu V, Manem VS, Ratanasirigulchai N, Chen GM, et al. MetaGxData: Clinically annotated breast, ovarian and pancreatic cancer datasets and their use in generating a multi-cancer gene signature. Scientific Reports. 2019;9(1):8770.

74. Ramos M. CuratedTCGAData: Curated data from the cancer genome atlas (tcga) as multiassayexperiment objects. 2019.

75. Veer LJ van't, Paik S, Hayes DF. Gene Expression Profiling of Breast Cancer: A New Tumor Marker. Journal of Clinical Oncology [Internet]. 2005 Mar;23(8):1631–5. Available from: http://www.ncbi.nlm.nih.gov/pubmed/15755970%20http://ascopubs.org/doi/10.1200/JCO.2005.12.005

76. Bao T, Davidson NE. Gene expression profiling of breast cancer. Advances in surgery [Internet]. 2008;42:249–60. Available from: http://www.ncbi.nlm.nih.gov/pubmed/18953822%20http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2775529

77. Bumgarner R. Overview of DNA microarrays: types, applications, and their future. Current protocols in molecular biology [Internet]. 2013 Jan;Chapter 22:Unit 22.1. Available from: http://www.ncbi.nlm.nih.gov/pubmed/23288464%20http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4011503

78. BeadArray microarray technology, https://www.illumina.com/science/technology/beadarray-technology.html.

79. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nature Reviews Genetics [Internet]. 2009 Jan;10(1):57–63. Available from: http://www.nature.com/doifinder/10.1038/nrg2484

80. Goossens N, Nakagawa S, Sun X, Hoshida Y. Cancer biomarker discovery and validation. Translational cancer research [Internet]. 2015 Jun;4(3):256–69. Available from: http://www.ncbi.nlm.nih.gov/pubmed/26213686%20http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4511498

81. McVeigh TP, Kerin MJ. Clinical use of the Oncotype DX genomic test to guide treatment decisions for patients with invasive breast cancer. Breast cancer (Dove Medical Press) [Internet]. 2017;9:393–400. Available from: http://www.ncbi.nlm.nih.gov/pubmed/28615971%20http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5459968

82. Cronin M, Sangli C, Liu M-L, Pho M, Dutta D, Nguyen A, et al. Analytical Validation of the Oncotype DX Genomic Diagnostic Test for Recurrence Prognosis and Therapeutic Response Prediction in Node-Negative, Estrogen Receptor-Positive Breast Cancer. Clinical Chemistry [Internet]. 2007 Apr;53(6):1084–91. Available from: http://www.ncbi.nlm

.nih.gov/pubmed/17463177%20http://www.clinchem.org/cgi/doi/10.1373/clinchem.2006.076497

83. Paik S, Tang G, Shak S, Kim C, Baker J, Kim W, et al. Gene Expression and Benefit of Chemotherapy in Women With Node-Negative, Estrogen Receptor?Positive Breast Cancer. Journal of Clinical Oncology [Internet]. 2006 Aug;24(23):3726–34. Available from: http://www.ncbi.nlm.nih.gov/pubmed/16720680%20http://ascopubs.org/doi/10.1200/JCO.2005.04.7985

84. Beumer IJ, Persoon M, Witteveen A, Dreezen C, Chin S-F, Sammut S-J, et al. Prognostic Value of MammaPrint() in Invasive Lobular Breast Cancer. Biomarker insights [Internet]. 2016;11:139–46. Available from: http://www.ncbi.nlm.nih.gov/pubmed/27980389%20http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5153320

85. Cardoso F, Veer LJ van't, Bogaerts J, Slaets L, Viale G, Delaloge S, et al. 70-Gene Signature as an Aid to Treatment Decisions in Early-Stage Breast Cancer. New England Journal of Medicine [Internet]. 2016 Aug;375(8):717–29. Available from: http://www.nejm.org/doi/10.1056/NEJMoa1602253

86. Drukker CA, Bueno-de-Mesquita JM, Retèl VP, Harten WH van, Tinteren H van, Wesseling J, et al. A prospective evaluation of a breast cancer prognosis signature in the observational RASTER study. International journal of cancer [Internet]. 2013 Aug;133(4):929–36. Available from: http://www.ncbi.nlm.nih.gov/pubmed/23371464%20http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3734625

87. Gy?rffy B, Hatzis C, Sanft T, Hofstatter E, Aktas B, Pusztai L. Multigene prognostic tests in breast cancer: past, present, future. Breast cancer research : BCR [Internet]. 2015 Jan;17(1):11. Available from: http://www.ncbi.nlm.nih.gov/pubmed/25848861%20http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4307898

88. Dowsett M, Sestak I, Lopez-Knowles E, Sidhu K, Dunbier AK, Cowens JW, et al. Comparison of PAM50 Risk of Recurrence Score With Onco type DX and IHC4 for Predicting Risk of Distant Recurrence After Endocrine Therapy. Journal of Clinical Oncology [Internet]. 2013 Aug;31(22):2783–90. Available from: http://www.ncbi.nlm.nih.gov/pubmed/23816962%20http://ascopubs.org/doi/10.1200/JCO.2012.46.1558

89. Filho OM, Ignatiadis M, Sotiriou C. Genomic Grade Index: An important tool for assessing breast cancer tumor grade and prognosis. Critical Reviews in Oncology/Hematology [Internet]. 2011 Jan;77(1):20–9. Available from: http://www.ncbi.nlm.nih.gov/pubmed/20138540%20http://linkinghub.elsevier.com/retrieve/pii/S1040842810000120

90. Müller BM, Keil E, Lehmann A, Winzer K-J, Richter-Ehrenstein C, Prinzler J, et al. The EndoPredict Gene-Expression Assay in Clinical Practice - Performance and Impact on Clinical Decisions. Lo AW, editor. PLoS ONE [Internet]. 2013 Jun;8(6):e68252. Available from: http://dx.plos.org/10.1371/journal.pone.0068252

91. Zhang Y, Schnabel CA, Schroeder BE, Jerevall P-L, Jankowitz RC, Fornander T, et al. Breast cancer index identifies early-stage estrogen receptor-positive breast cancer patients at risk for early- and late-distant recurrence. Clinical cancer research : an official journal of the American Association for Cancer Research [Internet]. 2013 Aug;19(15):4196–205. Available from: http://www.ncbi.nlm.nih.gov/pubmed/23757354

92. Lanigan F, Brien GL, Fan Y, Madden SF, Jerman E, Maratha A, et al. Delineating

transcriptional networks of prognostic gene signatures refines treatment recommendations for lymph node-negative breast cancer patients. FEBS Journal [Internet]. 2015 Sep;282(18):3455–73. Available from: http://doi.wiley.com/10.1111/febs.13354

93. Untch M, Konecny GE, Paepke S, Minckwitz G von. Current and future role of neoadjuvant therapy for breast cancer. The Breast [Internet]. 2014 Oct;23(5):526–37. Available from: http://www.ncbi.nlm.nih.gov/pubmed/25034931%20http://linkinghub.elsevier.com/retrieve/pii/S0960977614001143

94. Macaskill EJ, Dixon JM. Neoadjuvant Use of Endocrine Therapy in Breast Cancer. The Breast Journal [Internet]. 2007 May;13(3):243–50. Available from: http://www.ncbi.nlm.nih.gov/pubmed/17461898%20http://doi.wiley.com/10.1111/j.1524-4741.2007.00417.x

95. Sims AH, Smethurst GJ, Hey Y, Okoniewski MJ, Pepper SD, Howell A, et al. The removal of multiplicative, systematic bias allows integration of breast cancer gene expression datasets ? improving meta-analysis and prediction of prognosis. BMC Medical Genomics [Internet]. 2008 Dec;1(1):42. Available from: http://bmcmedgenomics.biomedcentral.com/articles/10.1186/1755-8794-1-42

96. Pearce DA, Arthur LM, Turnbull AK, Renshaw L, Sabine VS, Thomas JS, et al. Tumour sampling method can significantly influence gene expression profiles derived from neoadjuvant window studies. Scientific Reports [Internet]. 2016 Sep;6(1):29434. Available from: http://www.nature.com/articles/srep29434

97. Urruticoechea A, Smith IE, Dowsett M. Proliferation marker ki-67 in early breast cancer. Journal of clinical oncology. 2005;23(28):7212–20.

98. Makris A, Powles T, Dowsett M, Osborne C, Trott P, Fernando I, et al. Prediction of response to neoadjuvant chemoendocrine therapy in primary breast carcinomas. Clinical Cancer Research. 1997;3(4):593–600.

99. Faneyte IF, Schrama JG, Peterse JL, Remijnse PL, Rodenhuis S, Van de Vijver M. Breast cancer response to neoadjuvant chemotherapy: Predictive markers and relation with outcome. British journal of Cancer. 2003;88(3):406.

100. Gee JMW, Hutcheson IR. Understanding endocrine resistance: the critical need for sequential samples from clinical breast cancer and novel in vitro models. Breast cancer research : BCR [Internet]. 2005;7(5):187–9. Available from: http://www.ncbi.nlm.nih.gov/pubmed/16168136%20http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC1242147

101. Gutierrez MC, Detre S, Johnston S, Mohsin SK, Shou J, Allred DC, et al. Molecular Changes in Tamoxifen-Resistant Breast Cancer: Relationship Between Estrogen Receptor, HER-2, and p38 Mitogen-Activated Protein Kinase. Journal of Clinical Oncology [Internet]. 2005 Apr;23(11):2469–76. Available from: http://www.ncbi.nlm.nih.gov/pubmed/15753463%20http://ascopubs.org/doi/10.1200/JCO.2005.01.172

102. Turnbull AK, Arthur LM, Renshaw L, Larionov AA, Kay C, Dunbier AK, et al. Accurate Prediction and Validation of Response to Endocrine Therapy in Breast Cancer. Journal of Clinical Oncology [Internet]. 2015 Jul;33(20):2270–8. Available from: http://www.ncbi.nlm.nih.gov/pubmed/26033813%20http://ascopubs.org/doi/10.1200/JCO.2014.57.8963

103. AK Turnbull LR Anu Fernando. EA2Clin: A Novel Immunohitochemical Prognostic and Predictive Test for Patients with Estrogen Receptor PositiveBreast Cancer.

104. Miller WR, Larionov A, Renshaw L, Anderson TJ, Walker JR, Krause A, et al. Gene Expression Profiles Differentiating Between Breast Cancers Clinically Responsive or Resistant to Letrozole. Journal of Clinical Oncology [Internet]. 2009 Mar;27(9):1382–7. Available from: http://www.ncbi.nlm.nih.gov/pubmed/19224856%20http://ascopubs.org/doi/10.1200/JCO.2008.16.8849

105. Dunbier AK, Ghazoui Z, Anderson H, Salter J, Nerurkar A, Osin P, et al. Molecular Profiling of Aromatase Inhibitor-Treated Postmenopausal Breast Tumors Identifies Immune-Related Correlates of Resistance. Clinical Cancer Research [Internet]. 2013 May;19(10):2775–86. Available from: http://www.ncbi.nlm.nih.gov/pubmed/23493347%20http://clincancerres.aacrjournals.org/cgi/doi/10.1158/1078-0432.CCR-12-1000

106. Morrogh M, Andrade VP, Patil AJ, Qin L-X, Mo Q, Sakr R, et al. Differentially Expressed Genes in Window Trials are Influenced by the Wound-Healing Process: Lessons Learned from a Pilot Study with Anastrozole. Journal of Surgical Research [Internet]. 2012 Jul;176(1):121–32. Available from: http://www.ncbi.nlm.nih.gov/pubmed/21777924%20http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4509686%20http://linkinghub.elsevier.com/retrieve/pii/S0022480411005166

107. Brandão RD, Veeck J, Van de Vijver KK, Lindsey P, Vries B de, Elssen CH van, et al. A randomised controlled phase II trial of pre-operative celecoxib treatment reveals anti-tumour transcriptional response in primary breast cancer. Breast Cancer Research [Internet]. 2013 Apr;15(2):R29. Available from: http://www.ncbi.nlm.nih.gov/pubmed/23566419%20http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3672758%20http://breast-cancer-research.biomedcentral.com/articles/10.1186/bcr3409

108. Mackay A, Urruticoechea A, Dixon JM, Dexter T, Fenwick K, Ashworth A, et al. Molecular response to aromatase inhibitor treatment in primary breast cancer. Breast Cancer Research [Internet]. 2007 Jun;9(3):R37. Available from: http://www.ncbi.nlm.nih.gov/pubmed/17555561%20http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC1929101%20http://breast-cancer-research.biomedcentral.com/articles/10.1186/bcr1732

109. Sabine VS, Sims AH, Macaskill EJ, Renshaw L, Thomas JS, Dixon JM, et al. Gene expression profiling of response to mTOR inhibitor everolimus in pre-operatively treated post-menopausal women with oestrogen receptor-positive breast cancer. Breast Cancer Research and Treatment [Internet]. 2010 Jul;122(2):419–28. Available from: http://www.ncbi.nlm.nih.gov/pubmed/20480226%20http://link.springer.com/10.1007/s10549-010-0928-6

110. Korde LA, Lusa L, McShane L, Lebowitz PF, Lukes L, Camphausen K, et al. Gene expression pathway analysis to predict response to neoadjuvant docetaxel and capecitabine for breast cancer. Breast Cancer Research and Treatment [Internet]. 2010 Feb;119(3):685–99. Available from: http://www.ncbi.nlm.nih.gov/pubmed/20012355%20http://link.springer.com/10.1007/s10549-009-0651-3

111. Tempfer C, Pils D, Klar M, Orlowsk-Volk M, Zur Hausen A, Jäger M, et al. Basal-like molecular subtype and HER4 up-regulation and response to neoadjuvant chemotherapy in breast cancer. Oncology Reports [Internet]. 2011 Jul;26(4):1037–45. Available from: http://www.ncbi.nlm.nih.gov/pubmed/21769435%20http://www.spandidos-publications.com/10.3892/or.2011.1392

112. Hannemann J, Oosterkamp HM, Bosch CAJ, Velds A, Wessels LFA, Loo C, et

al. Changes in Gene Expression Associated With Response to Neoadjuvant Chemotherapy in Breast Cancer. Journal of Clinical Oncology [Internet]. 2005 May;23(15):3331–42. Available from: http://www.ncbi.nlm.nih.gov/pubmed/15908647%20http://ascopubs.org/doi/10.1200/JCO.2005.09.077

113. Chen C, Grennan K, Badner J, Zhang D, Gershon E, Jin L, et al. Removing Batch Effects in Analysis of Expression Microarray Data: An Evaluation of Six Batch Adjustment Methods. Kliebenstein D, editor. PLoS ONE [Internet]. 2011 Feb;6(2):e17238. Available from: http://dx.plos.org/10.1371/journal.pone.0017238

114. Müller C, Schillert A, Röthemeier C, Trégouët D-A, Proust C, Binder H, et al. Removing Batch Effects from Longitudinal Gene Expression - Quantile Normalization Plus ComBat as Best Approach for Microarray Transcriptome Data. Kaderali L, editor. PLOS ONE [Internet]. 2016 Jun;11(6):e0156594. Available from: http://dx.plos.org/10.1371/journal.pone.0156594

115. Moir N, Sims A. Compositional differences in tnbc presented in meta-analysis of expression profile datasets. Unpublished. 2020;

116. Saini A, Hou J, Zhou W. Breast cancer prognosis risk estimation using integrated gene expression and clinical data. BioMed research international [Internet]. 2014;2014:459203. Available from: http://www.ncbi.nlm.nih.gov/pubmed/24949450%20http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4052785

117. Shen R, Ghosh D, Chinnaiyan AM. Prognostic meta-signature of breast cancer developed by two-stage mixture modeling of microarray data. BMC genomics [Internet]. 2004 Dec;5(1):94. Available from: http://www.ncbi.nlm.nih.gov/pubmed/15598354%20http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC544889

118. Teschendorff AE, Naderi A, Barbosa-Morais NL, Pinder SE, Ellis IO, Aparicio S, et al. A consensus prognostic gene expression classifier for ER positive breast cancer. Genome biology [Internet]. 2006;7(10):R101. Available from: http://www.ncbi.nlm.nih.gov/pubmed/17076897%20http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC1794561

119. Simon R. Development and Validation of Therapeutically Relevant Multi-Gene Biomarker Classifiers. JNCI: Journal of the National Cancer Institute [Internet]. 2005 Jun;97(12):866–7. Available from: http://academic.oup.com/jnci/article/97/12/866/2544080/Development-and-Validation-of-Therapeutically

120. Ein-Dor L, Zuk O, Domany E. Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer. Proceedings of the National Academy of Sciences [Internet]. 2006 Apr;103(15):5923–8. Available from: http://www.ncbi.nlm.nih.gov/pubmed/16585533%20http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC1458674%20http://www.pnas.org/cgi/doi/10.1073/pnas.0601231103

121. Naderi A, Teschendorff AE, Barbosa-Morais NL, Pinder SE, Green AR, Powe DG, et al. A gene-expression signature to predict survival in breast cancer across independent data sets. Oncogene [Internet]. 2007 Mar;26(10):1507–16. Available from: http://www.ncbi.nlm.nih.gov/pubmed/16936776%20http://www.nature.com/doifinder/10.1038/sj.onc.1209920

122. Hamid JS, Hu P, Roslin NM, Ling V, Greenwood CMT, Beyene J. Data integration in genetics and genomics: methods and challenges. Human genomics and proteomics : HGP [Internet]. 2009 Jan;2009. Available from: http://www.ncbi.nlm.nih.gov/pubmed/

20948564%20http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2950414

123. Choi JK, Yu U, Kim S, Yoo OJ. Combining multiple microarray studies and modeling interstudy variation. Bioinformatics (Oxford, England) [Internet]. 2003;19 Suppl 1:i84–90. Available from: http://www.ncbi.nlm.nih.gov/pubmed/12855442

124. Shabalin AA, Tjelmeland H, Fan C, Perou CM, Nobel AB. Merging two gene-expression studies via cross-platform normalization. Bioinformatics [Internet]. 2008 May;24(9):1154–60. Available from: http://www.ncbi.nlm.nih.gov/pubmed/18325927%20https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btn083

125. Paquet ER, Hallett MT. Absolute Assignment of Breast Cancer Intrinsic Molecular Subtype. JNCI: Journal of the National Cancer Institute [Internet]. 2015 Jan;107(1). Available from: https://academic.oup.com/jnci/article-lookup/doi/10.1093/jnci/dju357

126. Dowsett M, Smith I, Robertson J, Robison L, Pinhel I, Johnson L, et al. Endocrine therapy, new biologicals, and new study designs for presurgical studies in breast cancer. Journal of the National Cancer Institute Monographs. 2011;2011(43):120–3.

127. Bliss J, Robison L, Webster-Smith M, Emson M, Kilburn L, Smith I, et al. OT2-03-04: A trial model for the future in the search for personalised medicine–the uk poetic and ephos-b perioperative trials experience. AACR; 2011.

128. Smith I, Johnson L, Dowsett M, Robertson J, Robison L, Kokan J, et al. Trial of perioperative endocrine therapy: Individualizing care (poetic). Journal of Clinical Oncology. 2011;29(15_suppl):TPS117–7.

129. Smith IE, Dowsett M, Ebbs SR, Dixon JM, Skene A, Blohmer J, et al. Neoadjuvant treatment of postmenopausal breast cancer with anastrozole, tamoxifen, or both in combination: The immediate preoperative anastrozole, tamoxifen, or combined with tamoxifen (impact) multicenter double-blind randomized trial. J Clin Oncol. 2005;23(22):5108–16.

130. Ellis MJ, Suman VJ, Hoog J, Goncalves R, Sanati S, Creighton CJ, et al. Ki67 Proliferation Index as a Tool for Chemotherapy Decisions During and After Neoadjuvant Aromatase Inhibitor Treatment of Breast Cancer: Results From the American College of Surgeons Oncology Group Z1031 Trial (Alliance). Journal of clinical oncology : official journal of the American Society of Clinical Oncology [Internet]. 2017 Apr;35(10):1061–9. Available from: http://ascopubs.org/doi/10.1200/JCO.2016.69.4406%20http://www.ncbi.nlm.nih.gov/pubmed/28045625%20http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5455353

131. Schott AF, Hayes DF. Defining the benefits of neoadjuvant chemotherapy for breast cancer. Journal of clinical oncology: official journal of the American Society of Clinical Oncology. 2012;30(15):1747.

132. Esserman LJ, Berry DA, DeMichele A, Carey L, Davis SE, Buxton M, et al. Pathologic complete response predicts recurrence-free survival more effectively by cancer subset: Results from the i-spy 1 trial—calgb 150007/150012, acrin 6657. Journal of Clinical Oncology. 2012;30(26):3242.

133. Untch M, Konecny GE, Paepke S, Minckwitz G von. Current and future role of neoadjuvant therapy for breast cancer. The Breast. 2014;23(5):526–37.

134. Asselain B, Barlow W, Bartlett J, Bergh J, Bergsten-Nordström E, Bliss J, et al.

Long-term outcomes for neoadjuvant versus adjuvant chemotherapy in early breast cancer: Meta-analysis of individual patient data from ten randomised trials. The Lancet Oncology. 2018;19(1):27–39.

135. Sims AH, Bartlett JM. Approaches towards expression profiling the response to treatment. BioMed Central; 2008.

136. Chia YH, Ellis MJ, Ma CX. Neoadjuvant endocrine therapy in primary breast cancer: Indications and use as a research tool. British Journal of Cancer [Internet]. 2010 Aug;103(6):759–64. Available from: http://dx.doi.org/10.1038/sj.bjc.6605845

137. Turnbull AK, Arthur LM, Renshaw L, Larionov AA, Kay C, Dunbier AK, et al. Accurate prediction and validation of response to endocrine therapy in breast cancer. Journal of Clinical Oncology. 2015;33(20):2270–8.

138. Ellis MJ, Suman VJ, Hoog J, Goncalves R, Sanati S, Creighton CJ, et al. Ki67 proliferation index as a tool for chemotherapy decisions during and after neoadjuvant aromatase inhibitor treatment of breast cancer: Results from the american college of surgeons oncology group z1031 trial (alliance). Journal of Clinical Oncology. 2017;35(10):1061.

139. Sørlie T, Perou CM, Fan C, Geisler S, Aas T, Nobel A, et al. Gene expression profiles do not consistently predict the clinical treatment response in locally advanced breast cancer. Molecular cancer therapeutics. 2006;5(11):2914–8.

140. Hatzis C, Pusztai L, Valero V, Booser DJ, Esserman L, Lluch A, et al. A genomic predictor of response and survival following taxane-anthracycline chemotherapy for invasive breast cancer. Jama. 2011;305(18):1873–81.

141. Magbanua MJM, Wolf DM, Yau C, Davis SE, Crothers J, Au A, et al. Serial expression analysis of breast tumors during neoadjuvant chemotherapy reveals changes in cell cycle and immune pathways associated with recurrence and response. Breast Cancer Research. 2015;17(1):73.

142. Hannemann J, Oosterkamp HM, Bosch C, Velds A, Wessels L, Loo C, et al. Changes in gene expression associated with response to neoadjuvant chemotherapy in breast cancer. J Clin Oncol. 2005;23(15):3331–42.

143. Gonzalez-Angulo AM, Iwamoto T, Liu S, Chen H, Do K-A, Hortobagyi GN, et al. Gene expression, molecular class changes, and pathway analysis after neoadjuvant systemic therapy for breast cancer. Clinical Cancer Research. 2012;18(4):1109–19.

144. Smyth GK, Michaud J, Scott HS. Use of within-array replicate spots for assessing differential expression in microarray experiments. Bioinformatics. 2005;21(9):2067–75.

145. R Core Team. R: A language and environment for statistical computing [Internet]. Vienna, Austria: R Foundation for Statistical Computing; 2019. Available from: https://www.R-project.org/

146. Huber W, Carey VJ, Gentleman R, Anders S, Carlson M, Carvalho BS, et al. Orchestrating high-throughput genomic analysis with Bioconductor. Nature Methods [Internet]. 2015;12(2):115–21. Available from: http://www.nature.com/nmeth/journal/v12/n2/full/nmeth.3252.html

147. Howe E, Holton K, Nair S, Schlauch D, Sinha R, Quackenbush J. Mev: Multiexperiment viewer. In: Biomedical informatics for cancer research. Springer; 2010. pp. 267–77.

148. Sherman BT, Tan Q, Collins JR, Alvord WG, Roayaei J, Stephens R, et al. The david gene functional classification tool: A novel biological module-centric algorithm to functionally analyze large gene lists. Genome biology. 2007;8(9):R183.

149. Abraham A, Pedregosa F, Eickenberg M, Gervais P, Mueller A, Kossaifi J, et al. Machine learning for neuroimaging with scikit-learn. Frontiers in neuroinformatics. 2014;8:14.

150. Tang Y, Li W. Lfda: An r package for local fisher discriminant analysis and visualization. arXiv preprint arXiv:161209219. 2016;

151. Tusher VG, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. Proceedings of the National Academy of Sciences. 2001;98(9):5116–21.

152. Bewick V, Cheek L, Ball J. Statistics review 12: Survival analysis. Critical care. 2004;8(5):389.

153. Gendoo DM, Ratanasirigulchai N, Schröder MS, Paré L, Parker JS, Prat A, et al. Genefu: An r/bioconductor package for computation of gene expression-based signatures in breast cancer. Bioinformatics. 2015;32(7):1097–9.

154. Sorlie T, Tibshirani R, Parker J, Hastie T, Marron J, Nobel A, et al. Repeated observation of breast tumor subtypes in independent gene expression data sets. Proceedings of the National Academy of Sciences of the United States of America. 2003;100(14):8418–23.

155. Hu Z, Fan C, Oh DS, Marron J, He X, Qaqish BF, et al. The molecular portraits of breast tumors are conserved across microarray platforms. BMC genomics. 2006;7(1):96.

156. Parker JS, Mullins M, Cheang MC, Leung S, Voduc D, Vickery T, et al. Supervised risk predictor of breast cancer based on intrinsic subtypes. Journal of clinical oncology. 2009;27(8):1160.

157. Desmedt C, Haibe-Kains B, Wirapati P, Buyse M, Larsimont D, Bontempi G, et al. Biological processes associated with breast cancer clinical outcome depend on the molecular subtypes. Clinical cancer research. 2008;14(16):5158–65.

158. Wirapati P, Sotiriou C, Kunkel S, Farmer P, Pradervand S, Haibe-Kains B, et al. Meta-analysis of gene expression profiles in breast cancer: Toward a unified understanding of breast cancer subtyping and prognosis signatures. Breast Cancer Research. 2008;10(4):R65.

159. Veer LJ van 't, Dai H, Vijver MJ van de, He YD, Hart AAM, Mao M, et al. Gene expression profiling predicts clinical outcome of breast cancer. Nature [Internet]. 2002 Jan;415(6871):530–6. Available from: http://dx.doi.org/10.1038/415530a

160. Van't Veer LJ, Dai H, Van De Vijver MJ, He YD, Hart AA, Mao M, et al. Gene expression profiling predicts clinical outcome of breast cancer. nature. 2002;415(6871):530.

161. Wallden B, Storhoff J, Nielsen T, Dowidar N, Schaper C, Ferree S, et al. Development and verification of the pam50-based prosigna breast cancer gene signature assay. BMC medical genomics. 2015;8(1):54.

162. Arthur LM, Turnbull AK, Webber VL, Larionov AA, Renshaw L, Kay C, et al. Molecular changes in lobular breast cancers in response to endocrine therapy. Cancer research. 2014;74(19):5371–6.

163. Tewari M, Krishnamurthy A, Shukla HS. Predictive markers of response to

neoadjuvant chemotherapy in breast cancer. Surgical oncology. 2008;17(4):301–11.

164. Colozza M, Azambuja E, Cardoso F, Sotiriou C, Larsimont D, Piccart M. Proliferative markers as prognostic and predictive tools in early breast cancer: Where are we now? Annals of oncology. 2005;16(11):1723–39.

165. Pohler E, Mamai O, Hirst J, Zamiri M, Horn H, Nomura T, et al. Haploinsufficiency for aagab causes clinically heterogeneous forms of punctate palmoplantar keratoderma. Nature genetics. 2012;44(11):1272.

166. Thul PJ, Åkesson L, Wiking M, Mahdessian D, Geladaki A, Blal HA, et al. A subcellular map of the human proteome. Science. 2017;356(6340).

167. Uhlen M, Zhang C, Lee S, Sjöstedt E, Fagerberg L, Bidkhori G, et al. A pathology atlas of the human cancer transcriptome. Science. 2017;357(6352).

168. Chatterjee G, Pai T, Hardiman T, Avery-Kiejda K, Scott RJ, Spencer J, et al. Molecular patterns of cancer colonisation in lymph nodes of breast cancer patients. Breast Cancer Research [Internet]. 2018 Nov;20(1). Available from: http://dx.doi.org/10.1186/s13058-018-1070-3

169. Escobar PF, Patrick RJ, Rybicki LA, Hicks D, Weng DE, Crowe JP. Prognostic significance of residual breast disease and axillary node involvement for patients who had primary induction chemotherapy for advanced breast cancer. Annals of surgical oncology. 2006;13(6):783–7.

170. Fisher B, Bauer M, Wickerham DL, Redmond CK, Fisher ER, Cruz AB, et al. Relation of number of positive axillary nodes to the prognosis of patients with primary breast cancer. An nsabp update. Cancer. 1983;52(9):1551–7.

171. Giuliano AE, McCall L, Beitsch P, Whitworth PW, Blumencranz P, Leitch AM, et al. Locoregional recurrence after sentinel lymph node dissection with or without axillary dissection in patients with sentinel lymph node metastases: The american college of surgeons oncology group z0011 randomized trial. Annals of surgery. 2010;252(3):426.

172. Nottegar A, Veronese N, Senthil M, Roumen R, Stubbs B, Choi A, et al. Extra-nodal extension of sentinel lymph node metastasis is a marker of poor prognosis in breast cancer patients: A systematic review and an exploratory meta-analysis. European Journal of Surgical Oncology (EJSO). 2016;42(7):919–25.

173. Li MH, Hou CL, Wang C, Sun AJ. HER-2, er, pr status concordance in primary breast cancer and corresponding metastatic lesion in lymph node in chinese women. Pathology-Research and Practice. 2016;212(4):252–7.

174. Tawfik K, Kimler BF, Davis MK, Fan F, Tawfik O. Ki-67 expression in axillary lymph node metastases in breast cancer is prognostically significant. Human pathology. 2013;44(1):39–46.

175. Zhao S, Xu L, Liu W, Lv C, Zhang K, Gao H, et al. Comparison of the expression of prognostic biomarkers between primary tumor and axillary lymph node metastases in breast cancer. International journal of clinical and experimental pathology. 2015;8(5):5744.

176. Falck A-K, Bendahl P-O, Chebil G, Olsson H, Fernö M, Rydén L. Biomarker expression and st gallen molecular subtype classification in primary tumours, synchronous lymph node metastases and asynchronous relapses in primary breast cancer patients with 10 years' follow-up. Breast cancer research and treatment. 2013;140(1):93–104.

177. Priedigkeit N, Hartmaier RJ, Chen Y, Vareslija D, Basudan A, Watters RJ, et al.

Intrinsic subtype switching and acquired erbb2/her2 amplifications and mutations in breast cancer brain metastases. JAMA oncology. 2017;3(5):666–71.

178. Lee JY, Park K, Lee E, Ahn T, Jung HH, Lim SH, et al. Gene expression profiling of breast cancer brain metastasis. Scientific reports. 2016;6:28623.

179. Wang Y, Klijn JG, Zhang Y, Sieuwerts AM, Look MP, Yang F, et al. Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. The Lancet. 2005;365(9460):671–9.

180. Paik S, Shak S, Tang G, Kim C, Baker J, Cronin M, et al. A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. New England Journal of Medicine. 2004;351(27):2817–26.

181. Sims AH. Bioinformatics and breast cancer: What can high-throughput genomic approaches actually tell us? Journal of Clinical Pathology [Internet]. 2009 Jan;62(10):879–85. Available from: http://dx.doi.org/10.1136/jcp.2008.060376

182. Shriver CD, Hueman MT, Ellsworth RE. Molecular signatures of lymph node status by intrinsic subtype: Gene expression analysis of primary breast tumors from patients with and without metastatic lymph nodes. Journal of Experimental & Clinical Cancer Research. 2014;33(1):116.

183. Lu X, Lu X, Wang ZC, Iglehart JD, Zhang X, Richardson AL. Predicting features of breast cancer with gene expression patterns. Breast cancer research and treatment. 2008;108(2):191.

184. Feng Y, Sun B, Li X, Zhang L, Niu Y, Xiao C, et al. Differentially expressed genes between primary cancer and paired lymph node metastases predict clinical outcome of node-positive breast cancer patients. Breast cancer research and treatment. 2007;103(3):319–29.

185. Suzuki M, Tarin D. Gene expression profiling of human lymph node metastases and matched primary breast carcinomas: Clinical implications. Molecular oncology. 2007;1(2):172–80.

186. Suzuki M, Tarin D. Gene expression profiling of human lymph node metastases and matched primary breast carcinomas: Clinical impl ications. Molecular Oncology [Internet]. 2007 Apr;1(2):172–80. Available from: http://dx.doi.org/10.1016/j.molonc.2007.03.005

187. Chatterjee G, Pai T, Hardiman T, Avery-Kiejda K, Scott RJ, Spencer J, et al. Molecular patterns of cancer colonisation in lymph nodes of breast cancer patients. Breast Cancer Research. 2018;20(1):143.

188. Vollebergh MA, Klijn C, Schouten PC, Wesseling J, Israeli D, Ylstra B, et al. Lack of genomic heterogeneity at high-resolution aCGH between primary breast cancers and their paired lymph node metastases. Diest PJ van, editor. PLoS ONE [Internet]. 2014 Aug;9(8):e103177. Available from: http://dx.doi.org/10.1371/journal.pone.0103177

189. Bownes RJ, Turnbull AK, Martinez-Perez C, Cameron DA, Sims AH, Oikonomidou O. On-treatment biomarkers can improve prediction of response to neoadjuvant chemotherapy in breast cancer. Breast Cancer Research. 2019;21(1):73.

190. Selli C, Turnbull AK, Pearce DA, Li A, Fernando A, Wills J, et al. Molecular changes during extended neoadjuvant letrozole treatment of breast cancer: Distinguishing acquired resistance from dormant tumours. Breast Cancer Research. 2019;21(1):2.

191. Ihaka R, Gentleman R. R: A language for data analysis and graphics. Journal of computational and graphical statistics. 1996;5(3):299–314.

192. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, et al. Bioconductor: Open software development for computational biology and bioinformatics. Genome biology. 2004;5(10):R80.

193. Breitling R, Armengaud P, Amtmann A, Herzyk P. Rank products: A simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments. FEBS letters. 2004;573(1-3):83–92.

194. Alexa A, Rahnenfuhrer J. TopGO: Enrichment analysis for gene ontology. 2019.

195. Wu D, Smyth GK. Camera: A competitive gene set test accounting for inter-gene correlation. Nucleic Acids Research [Internet]. 2012 May;40(17):e133–3. Available from: http://dx.doi.org/10.1093/nar/gks461

196. Barrett T, Suzek TO, Troup DB, Wilhite SE, Ngau W-C, Ledoux P, et al. NCBI geo: Mining millions of expression profiles—database and tools. Nucleic acids research. 2005;33(suppl_1):D562–6.

197. Cheng W-C, Shu W-Y, Li C-Y, Tsai M-L, Chang C-W, Chen C-R, et al. Intra- and inter-individual variance of gene expression in clinical studies. Jordan IK, editor. PLoS ONE [Internet]. 2012 Jun;7(6):e38650. Available from: http://dx.doi.org/10.1371/journal.pone.0038650

198. Pearce DA, Arthur LM, Turnbull AK, Renshaw L, Sabine VS, Thomas JS, et al. Tumour sampling method can significantly influence gene expression profiles derived from neoadjuvant window studies. Scientific Reports [Internet]. 2016 Sep;6(1):29434. Available from: http://www.nature.com/articles/srep29434

199. Hao X, Sun B, Hu L, Lähdesmäki H, Dunmire V, Feng Y, et al. Differential gene and protein expression in primary breast malignancies and their lymph node metastases as revealed by combined cDNA microarray and tissue microarray analysis. Cancer [Internet]. 2004 Mar;100(6):1110–22. Available from: http://dx.doi.org/10.1002/cncr.20095

200. Ellsworth RE, Seebach J, Field LA, Heckman C, Kane J, Hooke JA, et al. A gene expression signature that defines breast cancer metastases. Clinical & Experimental Metastasis [Internet]. 2008 Dec;26(3):205–13. Available from: http://dx.doi.org/10.1007/s10585-008-9232-9

201. Hao X, Sun B, Hu L, Lähdesmäki H, Dunmire V, Feng Y, et al. Differential gene and protein expression in primary breast malignancies and their lymph node metastases as revealed by combined cDNA microarray and tissue microarray analysis. Cancer: Interdisciplinary International Journal of the American Cancer Society. 2004;100(6):1110–22.

202. Weigelt B, Glas AM, Wessels LF, Witteveen AT, Peterse JL, Veer LJ van't. Gene expression profiles of primary breast tumors maintained in distant metastases. Proceedings of the National Academy of Sciences. 2003;100(26):15901–5.

203. Weigelt B, Wessels L, Bosma A, Glas A, Nuyten D, He Y, et al. No common denominator for breast cancer lymph node metastasis. British journal of cancer. 2005;93(8):924.

204. Weigelt B, Hu Z, He X, Livasy C, Carey LA, Ewend MG, et al. Molecular portraits

and 70-gene prognosis signature are preserved throughout the metastatic process of breast cancer. Cancer research. 2005;65(20):9155–8.

205. Lawler K, Papouli E, Naceur-Lombardelli C, Mera A, Ougham K, Tutt A, et al. Gene expression modules in primary breast cancers as risk factors for organotropic patterns of first metastatic spread: A case control study. Breast Cancer Research. 2017;19(1):113.

206. Calvo J, Sánchez-Cid L, Muñoz M, Lozano JJ, Thomson TM, Fernández PL. Infrequent loss of luminal differentiation in ductal breast cancer metastasis. PloS one. 2013;8(10):e78097.

207. Tobin NP, Lundberg A, Lindström LS, Harrell JC, Foukakis T, Carlsson L, et al. PAM50 provides prognostic information when applied to the lymph node metastases of advanced breast cancer patients. Clinical Cancer Research [Internet]. 2017 Sep;23(23):7225–31. Available from: http://dx.doi.org/10.1158/1078-0432.CCR-17-2301

208. Curtis C, Shah SP, Chin S-F, Turashvili G, Rueda OM, Dunning MJ, et al. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. Nature [Internet]. 2012 Apr;486(7403):346–52. Available from: http://dx.doi.org/10.1038/nature10983

209. Comprehensive molecular portraits of human breast tumours. Nature [Internet]. 2012 Sep;490(7418):61–70. Available from: http://dx.doi.org/10.1038/nature11412

210. Ramos M. CuratedTCGAData: Curated data from the cancer geneome atlas. 2019.

211. Toro-Domínguez D, Martorell-Marugán J, López-Domínguez R, García-Moreno A, González-Rumayor V, Alarcón-Riquelme ME, et al. ImaGEO: Integrative gene expression meta-analysis from geo database. Bioinformatics. 2018;35(5):880–2.

212. Forero DA. Available software for meta-analyses of genome-wide expression studies. PeerJ Preprints. 2019;7:e27708v1.

213. Cardoso F, Veer LJ van't, Bogaerts J, Slaets L, Viale G, Delaloge S, et al. 70-gene signature as an aid to treatment decisions in early-stage breast cancer. New England Journal of Medicine [Internet]. 2016 Aug;375(8):717–29. Available from: http://dx.doi.org/10.1056/NEJMoa1602253

214. Derks MGM, Velde CJH van de. Neoadjuvant chemotherapy in breast cancer: More than just downsizing. The Lancet Oncology [Internet]. 2018 Jan;19(1):2–3. Available from: http://dx.doi.org/10.1016/S1470-2045(17)30914-2

215. Bownes RJ, Turnbull AK, Martinez-Perez C, Cameron DA, Sims AH, Oikonomidou O. On-treatment biomarkers can improve prediction of response to neoadjuvant chemotherapy in breast cancer. Breast Cancer Research [Internet]. 2019 Jun;21(1). Available from: http://dx.doi.org/10.1186/s13058-019-1159-3

216. Turnbull AK, Arthur LM, Renshaw L, Larionov AA, Kay C, Dunbier AK, et al. Accurate prediction and validation of response to endocrine therapy in breast cancer. Journal of Clinical Oncology [Internet]. 2015 Jul;33(20):2270–8. Available from: http://dx.doi.org/10.1200/JCO.2014.57.8963

217. Turnbull AK, Kitchen RR, Larionov AA, Renshaw L, Dixon JM, Sims AH. Direct integration of intensity-level data from affymetrix and illumina microarrays improves statistical power for robust reanalysis. BMC medical genomics. 2012;5(1):35.

218. Leek JT, Storey JD. Capturing heterogeneity in gene expression studies by surrogate variable analysis. PLoS Genetics [Internet]. 2007;3(9):e161. Available from: http://dx.doi.org/10.1371/journal.pgen.0030161

219. Carvalho CM, Chang J, Lucas JE, Nevins JR, Wang Q, West M. High-dimensional sparse factor modeling: Applications in gene expression genomics. Journal of the American Statistical Association [Internet]. 2008 Dec;103(484):1438–56. Available from: http://dx.doi.org/10.1198/016214508000000869

220. Wang XV, Verhaak RGW, Purdom E, Spellman PT, Speed TP. Unifying gene expression measures from multiple platforms using factor analysis. Aerts S, editor. PLoS ONE [Internet]. 2011 Mar;6(3):e17691. Available from: http://dx.doi.org/10.1371/journal.pone.0017691

221. Kilpinen S, Autio R, Ojala K, Iljin K, Bucher E, Sara H, et al. Systematic bioinformatic analysis of expression levels of 17,330 human genes across 9,783 samples from 175 types of healthy and pathological tissues. Genome Biology [Internet]. 2008;9(9):R139. Available from: http://dx.doi.org/10.1186/gb-2008-9-9-r139

222. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical bayes methods. Biostatistics [Internet]. 2006 Apr;8(1):118–27. Available from: http://dx.doi.org/10.1093/biostatistics/kxj037

223. Müller C, Schillert A, Röthemeier C, Trégouët D-A, Proust C, Binder H, et al. Removing batch effects from longitudinal gene expression - quantile normalization plus combat as best approach for microarray transcriptome data. Kaderali L, editor. PLOS ONE [Internet]. 2016 Jun;11(6):e0156594. Available from: http://dx.doi.org/10.1371/journal.pone.0156594

224. Chen C, Grennan K, Badner J, Zhang D, Gershon E, Jin L, et al. Removing batch effects in analysis of expression microarray data: An evaluation of six batch adjustment methods. Kliebenstein D, editor. PLoS ONE [Internet]. 2011 Feb;6(2):e17238. Available from: http://dx.doi.org/10.1371/journal.pone.0017238

225. Kupfer P, Guthke R, Pohlers D, Huber R, Koczan D, Kinne RW. Batch correction of microarray data substantially improves the identification of genes differentially expressed in rheumatoid arthritis and osteoarthritis. BMC Medical Genomics [Internet]. 2012 Jun;5(1). Available from: http://dx.doi.org/10.1186/1755-8794-5-23

226. Taroni JN, Greene CS. Cross-platform normalization enables machine learning model training on microarray and rna-seq data simultaneously. BioRxiv. 2017;118349.

227. Davis S, Meltzer P. GEOquery: A bridge between the gene expression omnibus (geo) and bioconductor. Bioinformatics. 2007;14:1846–7.

228. Kitchen RR, Sabine VS, Simen AA, Dixon JM, Bartlett JM, Sims AH. Relative impact of key sources of systematic noise in affymetrix and illumina gene-expression microarray experiments. BMC Genomics [Internet]. 2011 Dec;12(1). Available from: http://dx.doi.org/10.1186/1471-2164-12-589

229. Huber W, Carey VJ, Gentleman R, Anders S, Carlson M, Carvalho BS, et al. Orchestrating high-throughput genomic analysis with Bioconductor. Nature Methods [Internet]. 2015;12(2):115–21. Available from: http://www.nature.com/nmeth/journal/v12/n2/full/nmeth.3252.html

230. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. limma powers differ-

ential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Research. 2015;43(7):e47.

231. Robinson MD, McCarthy DJ, Smyth GK. EdgeR: A bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010;26(1):139–40.

232. McCarthy DJ, Chen Y, Smyth GK. Differential expression analysis of multifactor rna-seq experiments with respect to biological variation. Nucleic Acids Research. 2012;40(10):4288–97.

233. Gendoo DMA, Ratanasirigulchai N, Schroeder MS, Pare L, Parker JS, Prat A, et al. Genefu: Computation of gene expression-based signatures in breast cancer [Internet]. 2019. Available from: http://www.pmgenomics.ca/bhklab/software/genefu

234. Wickham H. Ggplot2: Elegant graphics for data analysis [Internet]. Springer-Verlag New York; 2016. Available from: https://ggplot2.tidyverse.org

235. Wickham H. Tidyverse: Easily install and load the 'tidyverse' [Internet]. 2017. Available from: https://CRAN.R-project.org/package=tidyverse

236. Scharpf RB, Tjelmeland H, Parmigiani G, Nobel A. A bayesian model for cross-study differential gene expression. JASA [Internet]. 2009; Available from: 10.1198/jasa.2009.ap07611

237. Wickham H, Danenberg P, Eugster M. Roxygen2: In-line documentation for r [Internet]. 2018. Available from: https://CRAN.R-project.org/package=roxygen2

238. Law CW, Chen Y, Shi W, Smyth GK. Voom: Precision weights unlock linear model analysis tools for rna-seq read counts. Genome biology. 2014;15(2):R29.

239. Joseph A. CompareDF: Do a git style diff of the rows between two dataframes with similar structure [Internet]. 2019. Available from: https://CRAN.R-project.org/package=compareDF

240. Sims AH, Smethurst GJ, Hey Y, Okoniewski MJ, Pepper SD, Howell A, et al. The removal of multiplicative, systematic bias allows integration of breast cancer gene expression datasets–improving meta-analysis and prediction of prognosis. BMC medical genomics. 2008;1(1):42.

241. Larsen MJ, Thomassen M, Tan Q, Sørensen KP, Kruse TA. Microarray-based rna profiling of breast cancer: Batch effect removal improves cross-platform consistency. BioMed Research International [Internet]. 2014;2014:1–11. Available from: http://dx.doi.org/10.1155/2014/651751

242. Henry L, Wickham H. Purrr: Functional programming tools [Internet]. 2019. Available from: https://CRAN.R-project.org/package=purrr

243. Kolde R. Pheatmap: Pretty heatmaps [Internet]. 2019. Available from: https://CRAN.R-project.org/package=pheatmap

244. He H, Shen X. A ranked subspace learning method for gene expression data classification. In: IC-ai. 2007. pp. 358–64.

245. Al-Quraishi T, Abawajy JH, Chowdhury MU, Rajasegarar S, Abdalrada AS. Breast cancer recurrence prediction using random forest model. In: International conference on soft computing and data mining. Springer; 2018. pp. 318–29.

246. Scherer A. Batch effects and noise in microarray experiments: Sources and solutions. Vol. 868. John Wiley & Sons; 2009.

247. Chen C, Grennan K, Badner J, Zhang D, Gershon E, Jin L, et al. Removing batch

effects in analysis of expression microarray data: An evaluation of six batch adjustment methods. PloS one. 2011;6(2):e17238.

248. Stein CK, Qu P, Epstein J, Buros A, Rosenthal A, Crowley J, et al. Removing batch effects from purified plasma cell gene expression microarrays with modified combat. BMC bioinformatics. 2015;16(1):63.

249. Kupfer P, Guthke R, Pohlers D, Huber R, Koczan D, Kinne RW. Batch correction of microarray data substantially improves the identification of genes differentially expressed in rheumatoid arthritis and osteoarthritis. BMC medical genomics. 2012;5(1):23.

250. Eklund AC, Szallasi Z. Correction of technical bias in clinical microarray data improves concordance with known biological information. Genome biology. 2008;9(2):R26.

251. Engchuan W, Meechai A, Tongsima S, Chan JH. Handling batch effects on cross-platform classification of microarray data. International Journal of Advanced Intelligence Paradigms. 2016;8(1):59–76.

252. Müller C, Schillert A, Röthemeier C, Trégouët D-A, Proust C, Binder H, et al. Removing batch effects from longitudinal gene expression-quantile normalization plus combat as best approach for microarray transcriptome data. PloS one. 2016;11(6):e0156594.

253. Moretto M, Sonego P, Villaseñor-Altamirano AB, Engelen K. First step toward gene expression data integration: Transcriptomic data acquisition with command> _. BMC bioinformatics. 2019;20(1):54.

254. Ge SX, Son EW, Yao R. IDEP: An integrated web application for differential expression and pathway analysis of rna-seq data. BMC bioinformatics. 2018;19(1):534.

255. Jiang Y, Liang Y, Wang D, Xu D, Joshi T. A dynamic programing approach to integrate gene expression data and network information for pathway model generation. Kelso J, editor. Bioinformatics [Internet]. 2019 Jun; Available from: http://dx.doi.org/10.1093/bioinformatics/btz467

256. Grimes T, Potter SS, Datta S. Integrating gene regulatory pathways into differential network analysis of gene expression data. Scientific reports. 2019;9(1):5479.

257. Blazier AS, Papin JA. Integration of expression data in genome-scale metabolic network reconstructions. Frontiers in Physiology [Internet]. 2012;3. Available from: http://dx.doi.org/10.3389/fphys.2012.00299

258. Li Z, Gao N, Martini JWR, Simianer H. Integrating gene expression data into genomic prediction. Frontiers in Genetics [Internet]. 2019 Feb;10. Available from: http://dx.doi.org/10.3389/fgene.2019.00126

259. Moretto M, Sonego P, Villaseñor-Altamirano AB, Engelen K. First step toward gene expression data integration: Transcriptomic data acquisition with command> _. BMC bioinformatics. 2019;20(1):54.

260. Anders S, Huber W. Differential expression analysis for sequence count data. Genome Biology [Internet]. 2010;11:R106. Available from: http://genomebiology.com/2010/11/10/R106/

261. Del Carratore F, Jankevics A, Eisinga R, Heskes T, Hong F, Breitling R. RankProd 2.0: A refactored bioconductor package for detecting differentially expressed features in molecular profiling datasets. Bioinformatics. 2017;33(17):2774–5.

262. Morgan M, Falcon S, Gentleman R. GSEABase: Gene set enrichment data structures and methods. 2019.

263. Carsten Kuhl HT Ralf Tautenhahn. CAMERA:An integrated strategy for compound spectra extraction and annotation of liquid chromatography/mass spectrometry data sets. 2012.

264. Kanehisa M, Goto S. KEGG: Kyoto encyclopedia of genes and genomes. Nucleic acids research. 2000;28(1):27–30.

265. Van Rossum G, Drake Jr FL. Python tutorial. Centrum voor Wiskunde en Informatica Amsterdam, The Netherlands; 1995.

266. Scikit-learn: Machine learning in Python. Journal of Machine Learning Research. 2011;12:2825–30.

267. Cortes C, Vapnik V. Support-vector networks. Machine learning. 1995;20(3):273–97.

268. TensorFlow: Large-scale machine learning on heterogeneous systems [Internet]. 2015. Available from: http://tensorflow.org/

269. Chollet F, others. Keras. https://github.com/fchollet/keras; GitHub; 2015.

270. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. CoRR abs/1512.03385 (2015). 2015.

271. Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, et al. PyTorch: An imperative style, high-performance deep learning library. In: Wallach H, Larochelle H, Beygelzimer A, d'Alché-Buc F, Fox E, Garnett R, editors. Advances in neural information processing systems 32 [Internet]. Curran Associates, Inc. 2019. pp. 8024–35. Available from: http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf

272. Li W, Cerise JE, Yang Y, Han H. Application of t-sne to human genetic data. Journal of Bioinformatics and Computational Biology. 2017;15(04):1750017.

273. Buchholz TA, Stivers DN, Stec J, Ayers M, Clark E, Bolt A, et al. Global gene expression changes during neoadjuvant chemotherapy for human breast cancer. The Cancer Journal. 2002;8(6):461–8.

274. Modlich O, Prisack H-B, Munnes M, Audretsch W, Bojar H. Immediate gene expression changes after the first course of neoadjuvant chemotherapy in patients with primary breast cancer disease. Clinical cancer research. 2004;10(19):6418–31.

275. Dowsett M, Smith IE, Ebbs SR, Dixon JM, Skene A, Griffith C, et al. Short-term changes in ki-67 during neoadjuvant treatment of primary breast cancer with anastrozole or tamoxifen alone or combined correlate with recurrence-free survival. Clinical Cancer Research. 2005;11(2):951s–8s.

276. Dowsett M, Smith IE, Ebbs SR, Dixon JM, Skene A, Griffith C, et al. Proliferation and apoptosis as markers of benefit in neoadjuvant endocrine therapy of breast cancer. Clinical Cancer Research. 2006;12(3):1024s–30s.

277. Chang J, Ormerod M, Powles T, Allred D, Ashley S, Dowsett M. Apoptosis and proliferation as predictors of chemotherapy response in patients with breast carcinoma. Cancer: Interdisciplinary International Journal of the American Cancer Society. 2000;89(11):2145–52.

278. Juvekar A, Burga LN, Hu H, Lunsford EP, Ibrahim YH, Balmañà J, et al. Combining a pi3k inhibitor with a parp inhibitor provides an effective therapy for brca1-related

breast cancer. Cancer discovery. 2012;2(11):1048–63.

279. Costa RL, Han HS, Gradishar WJ. Targeting the pi3k/akt/mTOR pathway in triple-negative breast cancer: A review. Breast cancer research and treatment. 2018;169(3):397–406.

280. Eshlaghy AT, Poorebrahimi A, Ebrahimi M, Razavi AR, Ahmad LG. Using three machine learning techniques for predicting breast cancer recurrence. J Health Med Inform. 2013;4(2):124.

281. Khan S, Islam N, Jan Z, Din IU, Rodrigues JJC. A novel deep learning based framework for the detection and classification of breast cancer using transfer learning. Pattern Recognition Letters. 2019;125:1–6.

282. Turashvili G, Brogi E. Tumor heterogeneity in breast cancer. Frontiers in medicine. 2017;4:227.

283. Ellsworth RE, Blackburn HL, Shriver CD, Soon-Shiong P, Ellsworth DL. Molecular heterogeneity in breast cancer: State of the science and implications for patient care. In: Seminars in cell & developmental biology. Elsevier; 2017. pp. 65–72.

284. Sabine VS, Sims AH, Macaskill EJ, Renshaw L, Thomas JS, Dixon JM, et al. Gene expression profiling of response to mTOR inhibitor everolimus in pre-operatively treated post-menopausal women with oestrogen receptor-positive breast cancer. Breast cancer research and treatment. 2010;122(2):419–28.

285. Ein-Dor L, Kela I, Getz G, Givol D, Domany E. Outcome signature genes in breast cancer: Is there a unique set? Bioinformatics. 2004;21(2):171–8.

286. Venet D, Dumont JE, Detours V. Most random gene expression signatures are significantly associated with breast cancer outcome. PLoS computational biology. 2011;7(10):e1002240.

287. Ein-Dor L, Zuk O, Domany E. Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer. Proceedings of the National Academy of Sciences. 2006;103(15):5923–8.

288. Schrijver WA, Selenica P, Lee JY, Ng CK, Burke KA, Piscuoglio S, et al. Mutation profiling of key cancer genes in primary breast cancers and their distant metastases. Cancer research. 2018;78(12):3112–21.

289. Moreno F, Gayarre J, López-Tarruella S, Monte-Millán M del, Picornell AC, Álvarez E, et al. Concordance of genomic variants in matched primary breast cancer, metastatic tumor, and circulating tumor dna: The mirror study. JCO Precision Oncology. 2019;3:1–16.

290. Varešlija D, Priedigkeit N, Fagan A, Purcell S, Cosgrove N, O'Halloran PJ, et al. Transcriptome characterization of matched primary breast and brain metastatic tumors to detect novel actionable targets. JNCI: Journal of the National Cancer Institute. 2018;111(4):388–98.

291. Iwamoto T, Niikura N, Ogiya R, Yasojima H, Watanabe K-i, Kanbayashi C, et al. Distinct gene expression profiles between primary breast cancers and brain metastases from pair-matched samples. Scientific reports. 2019;9(1):1–8.

292. Nayak BK. Understanding the relevance of sample size calculation. Indian journal of ophthalmology. 2010;58(6):469.

293. Bendjilali N, MacLeon S, Kalra G, Willis SD, Hossian AN, Avery E, et al. Time-course analysis of gene expression during the saccharomyces cerevisiae hypoxic re-

sponse. G3: Genes, Genomes, Genetics. 2017;7(1):221–31.

294. Androulakis IP, Yang E, Almon RR. Analysis of time-series gene expression data: Methods, challenges, and opportunities. Annu Rev Biomed Eng. 2007;9:205–28.

295. Nygaard V, Rødland EA, Hovig E. Methods that remove batch effects while retaining group differences may lead to exaggerated confidence in downstream analyses. Biostatistics. 2016;17(1):29–39.

296. Tian T, Li X, Zhang J. MTOR signaling in cancer and mTOR inhibitors in solid tumor targeting therapy. International journal of molecular sciences. 2019;20(3):755.

297. Hynes NE, Boulay A. The mTOR pathway in breast cancer. Journal of mammary gland biology and neoplasia. 2006;11(1):53–61.

298. Steelman LS, Martelli AM, Cocco L, Libra M, Nicoletti F, Abrams SL, et al. The therapeutic potential of mTOR inhibitors in breast cancer. British journal of clinical pharmacology. 2016;82(5):1189–212.

299. Serra V, Scaltriti M, Prudkin L, Eichhorn PJ, Ibrahim YH, Chandarlapaty S, et al. PI3K inhibition results in enhanced her signaling and acquired erk dependency in her2-overexpressing breast cancer. Oncogene. 2011;30(22):2547.

300. Qin H, Liu L, Sun S, Zhang D, Sheng J, Li B, et al. The impact of pi3k inhibitors on breast cancer cell and its tumor microenvironment. PeerJ. 2018;6:e5092.

301. Yang J, Nie J, Ma X, Wei Y, Peng Y, Wei X. Targeting pi3k in cancer: Mechanisms and advances in clinical trials. Molecular cancer. 2019;18(1):26.

302. Berns K, Horlings HM, Hennessy BT, Madiredjo M, Hijmans EM, Beelen K, et al. A functional genetic approach identifies the pi3k pathway as a major determinant of trastuzumab resistance in breast cancer. Cancer cell. 2007;12(4):395–402.

303. Bedard PL, Hansen AR, Ratain MJ, Siu LL. Tumour heterogeneity in the clinic. Nature. 2013;501(7467):355–64.

304. Lüder Ripoli F, Mohr A, Conradine Hammer S, Willenbrock S, Hewicker-Trautwein M, Hennecke S, et al. A comparison of fresh frozen vs. Formalin-fixed, paraffin-embedded specimens of canine mammary tumors via branched-dna assay. International journal of molecular sciences. 2016;17(5):724.